Modeling the effect of blending multiple components on gasoline properties

Sandra Correa González

School of Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.
Espoo 30.9.2019

Supervisor

Prof. Annukka Santasalo-Aarnio

Advisors

MSc Anna Karvo

MSc Yuri Kroyan

Aalto University
School of Engineering

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Sandra Correa González

Title Modeling the effect of blending multiple components on gasoline properties

Degree programme Innovative and Sustainable Energy Engineering (ISEE)

Major Bioenergy **Code of major** ENG215

Supervisor Prof. Annukka Santasalo-Aarnio

Advisors MSc Anna Karvo, MSc Yuri Kroyan

Date 30.9.2019 **Number of pages** 80+4 **Language** English

Abstract

Global CO_2 emissions reached a new historical maximum in 2018 and transportation sector contributed to one fourth of those emissions. Road transport industry has started moving towards more sustainable solutions, however, market penetration for electric vehicles (EV) is still too slow while regulation for biofuels has become stricter due to the risk of inflated food prices and skepticism regarding their sustainability. In spite of this, Europe has ambitious targets for the next 30 years and impending strict policies resulting from these goals will definitely increase the pressure on the oil sector to move towards cleaner practices and products.

Although the use of biodiesel is quite extended and bioethanol is already used as a gasoline component, there are no alternative drop-in fuels compatible with spark ignition engines in the market yet. Alternative feedstock is widely available but its characteristics differ from those of crude oil, and lack of homogeneity and substantially lower availability complicate its integration in conventional refining processes. This work explores the possibility of implementing Machine Learning to develop predictive models for auto-ignition properties and to gain a better understanding of the blending behavior of the different molecules that conform commercial gasoline. Additionally, the methodology developed in this study aims to contribute to new characterization methods for conventional and renewable gasoline streams in a simpler, faster and more inexpensive way.

To build the models included in this thesis, a palette with seven different compounds was chosen: n-heptane, iso-octane, 1-hexene, cyclopentane, toluene, ethanol and ETBE. A data set containing 243 different combinations of the species in the palette was collected from literature, together with their experimentally measured RON and/or MON. Linear Regression based on Ordinary Least Squares was used as the baseline to compare the performance of more complex algorithms, namely Nearest Neighbors, Support Vector Machines, Decision Trees and Random Forest. The best predictions were obtained with a Support Vector Regression algorithm using a non-linear kernel, able to reproduce synergistic and antagonistic interaction between the seven molecules in the samples.

Keywords Gasoline, alternative feedstock, fuel blend, spark ignition engines, auto-ignition, RON, MON, Machine Learning, predictive models, Python

Preface

I want to thank my supervisor Professor Annukka Santasalo-Aarnio for her guidance and constant enthusiasm, my second supervisor Henry Persson for his significant contributions and my advisors Yuri Kroyan and Anna Karvo for their valuable comments and feedback along the process.

I also want to thank my family and friends for their unconditional trust and support.

Espoo, 30.9.2019

Sandra Correa González

Contents

List of Figures

List of Tables

Abbreviations

AKI	Antiknock Index
ANN	Artificial Neural Networks
BDC	Bottom Dead Center
CDU	Crude Distillation Unit
CFR	Cooperative Fuel Research
CWD	Cold Weather Driveability
DCU	Delayed coker unit
DT	Decision Trees
E_{100}	Volume percent evaporated at 100°C
E_{150}	Volume percent evaporated at 150°C
E_{70}	Volume percent evaporated at 70°C
E5	Gasoline blend containing 5% ethanol
E10	Gasoline blend containing 10% ethanol
E85	Gasoline blend containing 85% ethanol
ETBE	Ethyl tert-butyl ether
EV	Electric vehicle
FBPC	Fuel Blend Property Calculator
FCC	Fluidized catalytic cracking
FFV	Flexible-fuel vehicle
GHG	Greenhouse gases
HAGO	Heavy atmospheric gas oil
HCA	Hierarchical Cluster Analysis
HSR	Heavy straight-run
HTL	Hydrothermal liquefaction
HVGO	Heavy vacuum gas oil
k-NN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
LinSVR	Linear Support Vector Regression
LSR	Light straight-run
LTHR	Low-temperature heat release
MAE	Mean Absolute Error
MILP	Mixed integer linear programming
MINLP	Mixed integer non-linear programming
ML	Machine learning
MLP	Multi-layer Perceptron
MON	Motor Octane Number
MSE	Mean Squared Error
MTBE	Methyl tert-butyl ether
NDC	Nationally Determined Contributions
NTC	Negative-temperature coefficient
Nu-SVR	Nu-Support Vector Regression
OI	Octane Index
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PIONA	Paraffins, iso-paraffins, olefins, naphthenes, aromatics

PLS	Partial Least Squares
PRF	Primary Reference Fuels
QSPR	Quantitative Structure Property Relationship
RBF	Gaussian radial basis function
RF	Random Forest
RMSE	Root Mean Squared Error
r-NN	Radius-based Nearest Neighbors
RON	Research Octane Number
RVP	Reid vapor pressure
S	Octane sensitivity
SI	Spark ignition
SVM	Support Vector Machines
SVO	Straight vegetable oil
SVR	Epsilon-Support Vector Regression
TDC	Top Dead Center
TEL	Tetraethyllead
TLU	Threshold logic unit
TVP	True vapor pressure
VDU	Vacuum distillation unit

1 Introduction

After a period of stagnation, global CO_2 emissions raised again in 2017 and 2018, reaching a new historical maximum of 33.1 Gt. Improvements and deployment of clean and more efficient technologies have not been able to develop in parallel with an expanding global economy and growing energy demand [1]. Despite substantial investments on renewable energy sources, more than 80% of the world energy supply still derives from the combustion of fossil fuels [2] and contributes to the increasing concentration of greenhouse gases (GHG) in the atmosphere. Heat and electricity production is the main source of anthropogenic CO_2. Transportation sector comes after, responsible for roughly one fourth of the total emissions, where road traffic stands out from maritime transportation and aviation with 74% of those emissions [3]. The link between the increasing concentration of GHG in the atmosphere and climate change and global warming is widely supported by the scientific community [4]. By the end of 2018, atmospheric CO_2 levels were measured at 410 ppm [5] and the global average temperature was 0.82°C higher than it was in the late 19^{th} century [6].

The Kyoto protocol (1997) was the first international treaty where state parties committed to reduce their GHG emissions. Under this agreement, the European Union has developed specific targets and policies to reduced emissions from all sectors by 2020 [7]. At the Paris climate conference in 2015, 195 countries committed on a long-term agreement to keep global temperature rise well under 2°C from pre-industrial levels [8]. All the parties submitted their Nationally Determined Contributions (NDC) where each nation presented its strategy to reduce emissions levels and contribute to sustainable development. The European Union's NDC includes efforts for a 40% reduction of GHG emissions by 2030 [9]. This goal is part of a wider EU climate and energy framework that includes targets for renewable energy and energy efficiency, as well as additional targets for transportation. This new framework revises and updates the 2020 Energy Strategy, which aimed at 10% renewable penetration in the transport sector and 6% fuel decarbonization [10] and reflects on the Roadmap to a Single European Transport Area. By 2050, Europe aims to get cities free from fossil fuel powered vehicles, to increase aviation fuel sustainability and to cut transport related emissions by 60% compared to 1990 [11].

Despite increasing popularity of electric vehicles (EVs), their market penetration is still too slow. In 2017 only 1.5% of the vehicles sold in the EU were electric [12]. These figures reveal the need for biofuels to achieve decarbonization of the sector. Sustainable drop-in fuels fully compatible with the existing car fleet are the only way to cut down emissions in the short-term and give time for an EV infrastructure to develop. However, due to the risk of inflated food prices, the European Commission has set a cap of 7% biofuels coming from edible crops. In addition to that, all Member States must ensure 14% of the fuel supply for road and rail transport to derive from renewable sources within the next decade. [13] Therefore, large investments are now required to develop advanced biofuels derived from energy crops, waste and algae that can compete in the market with crude oil derived products.

In the described context, this master's thesis aims to contribute to increase the knowledge

on the interaction among gasoline components and to set the basis to be able to predict the properties of future green fuels. The scope of this paper is limited to alternative fuels and additives for spark ignition (SI) engines able to replace the extended use of fossil gasoline. Nevertheless, the tools investigated in this work are useful for the study and understanding of other alternative fuels. Likewise, the methodology in this thesis has been carefully developed to fit future purposes in similar areas of expertise.

1.1 Thesis structure

This master's thesis has been accomplished in the framework of DigiFuels project. DigiFuels is a collaborative project between Aalto University and Neste that seeks to enhance product development for SI engines. The project includes three different working packages, this study belonging to the first of them which pursues the development of a Fuel Blend Property Calculator (FBPC). This document, structured in 6 different sections, provides with a solid background for further research on the topic.

Chapter 1 shortly presents the global energy situation, with special attention to pollution levels and international efforts and policies to tackle climatic issues. Biofuels are introduced as an alternative to diminish GHG emission derived from the transportation sector.

Chapters 2 and 3 consist of an extensive literature study. Chapter 2 covers conventional and alternative gasoline-like fuels. Feedstock, properties and applications of different fuels are presented and compared. Moreover, it elaborates on the need of novel and more flexible tools to facilitate the inclusion of cleaner fuels and additives in the current scenario. Chapter 3 provides with a background on Machine Learning as a possible tool beyond traditional statistical analysis relevant for property modeling. In this context, supervised and unsupervised algorithms are discussed.

Chapter 4 explains the methodology adopted for this master's thesis. It reviews the data acquisition process and its post-processing. Moreover, it gives a detailed explanation of the training, validation and testing stages for the different algorithms explored in this thesis.

Chapter 5 presents and discusses the results of the proposed models. The performance of the models is assessed internally using training data and 10-fold cross-validation, as well as externally over a test set. These two approaches serve as a base for comparison of the different algorithms and to reflect back on the collected database.

Chapter 6 summarizes the work and highlights the main conclusion extracted from the previous chapters. Moreover, it reflects on the limitations of the models and difficulties encountered along the elaboration of this study, giving hints for possible future research areas.

1.2 Research questions

Overall, this master's thesis will try to give a response to the following questions:

1. How can variations in the raw materials affect chemical composition and final properties of gasoline?
2. Which novel tools can be used to predict future gasoline properties with changing composition and model the complex molecular interactions?
3. Are more advanced modeling tools able to reflect non-linear blending behaviors in gasoline?

2 Background

This section provides an overview of conventional and alternative fuels available for SI engines. Firstly, fossil-derived gasoline is presented together with its composition, properties and performance. Thereafter, biofuels and alternative feedstock are explored and the concept of renewable gasoline is introduced. Overall, this chapter aims to highlight the challenges in the transition towards sustainable raw materials and to evince the need of advanced tools to boost the deployment of cleaner fuels.

2.1 Gasoline

Gasoline is one of the multiple products derived from petroleum refining. High quality commercial fuels are the result of complex blending processes, where different light streams and additives are combined together to achieve the desired properties. As a result, over 600 or 700 different molecules can be present in a gasoline sample. Most of these molecules classify as small and light hydrocarbons with carbon number ranging between 4 and 12, although the exact chemical composition is widely variable. This variability depends on several factors such as blending strategy of the refinery, characteristics of the crude oil or regional and seasonal specification for the final products. [14, 15]

An overview of this chemical complexity of gasoline is given in Figure 2.1.

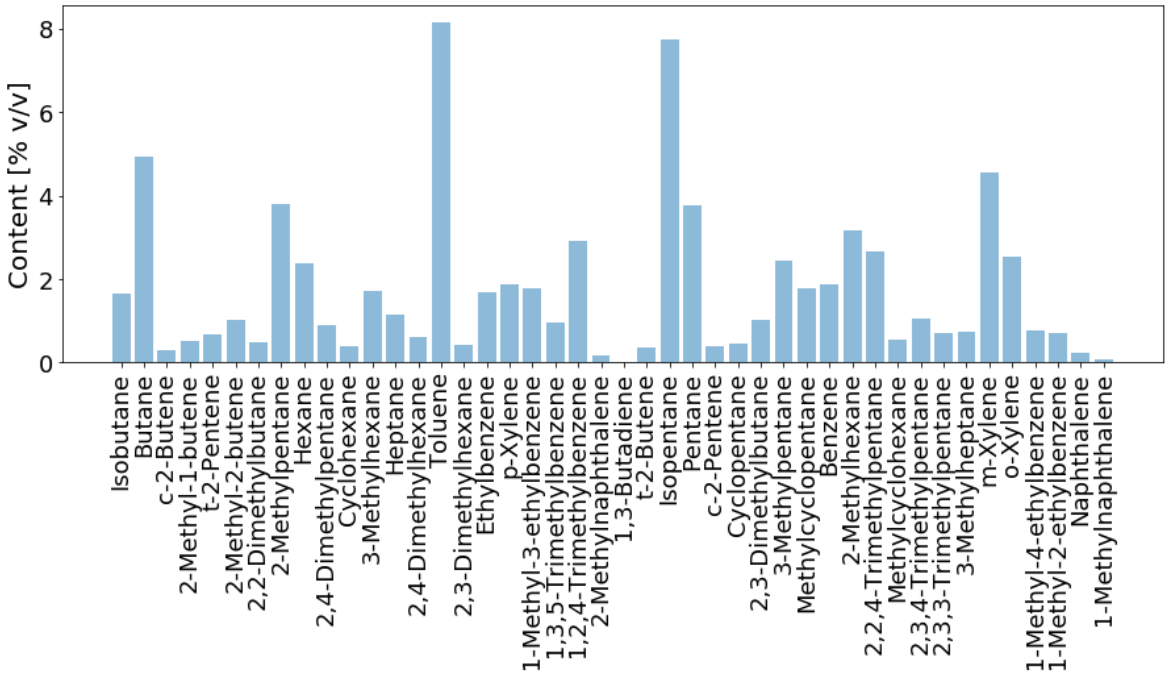

Figure 2.1 Frequent compounds found in commercial gasoline and their typical concentration [16]

Figure 2.1 shows the most abundant components in premium and regular unleaded gasolines collected over the course of one year in different regions of Canada [16]. A

total of 44 molecules are included, corresponding to compounds that may be found in commercial gasoline in a concentration grater than 1% v/v and which make up for 70 to 90% of the analyzed fuels. Molecules present in a higher concentration are expected to have a larger impact on the properties of the fuel, but non-linear blending interactions tend to lead to unpredictable results. Nevertheless, patterns and trends can be noticed within families of compounds. In the case of the oil industry, it is common to refer to the term PIONA which alludes to the classification of hydrocarbons under the following groups: n-paraffins (P), iso-paraffins (I), olefins (O), naphthenes (N) and aromatics (A). Figure 2.2 shows the typical distribution range of these five groups in commercial US gasolines, explained in more detail next.

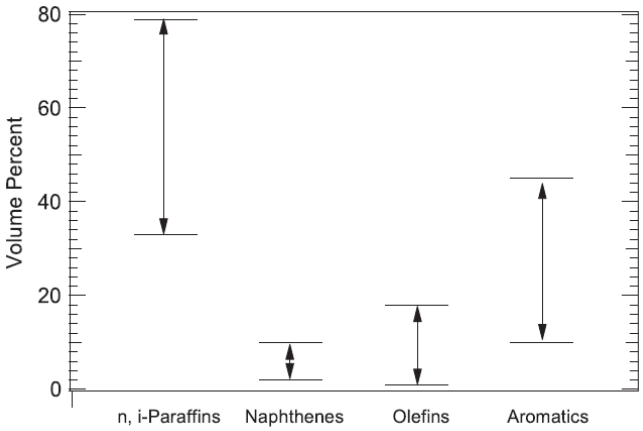

Figure 2.2 Typical PIONA composition ranges for US commercial gasoline [17]

- **Paraffins**, also known as alkanes, are open-chain saturated hydrocarbons with general chemical formula C_nH_{2n+2}. Paraffins can be subdivided into straight-chain or n-paraffins and branched-chain or iso-paraffins. Together they account for the highest fraction in gasoline, as shown in Figure 2.2, although the concentration of iso-paraffins is typically higher due to their higher octane numbers. [14, 17]

- **Olefins**, or alkenes, are unsaturated hydrocarbons. Unlike paraffins, olefins present double or triple carbon bonds which result in higher antiknock performance, but also in high octane sensitivity. Additionally, they have low oxidation stability and reduce storage lifespan of the fuel. Therefore, less than 20% olefins are blended into the final product, as shown in Figure 2.2 for the US case. [14, 17]

- **Aromatics** are unsaturated cyclic compounds that contain one or more benzene or similar ring structures [14]. Despite their good auto-ignition performance, an upper blending limit of 30-35% volume is set due to high particulate matter emissions associated with their combustion [17]. This is true for American gasolines as reflected by Figure 2.2, but also in Europe, where that limitation is set by the fuel standard EN 228 [18].

- **Naphthenes** are cyclic aliphatic compounds with general formula C_nH_{2n}. Generally, they present high boiling points and low octane numbers and they are prone to dehydrogenation into aromatics, hence their presence is kept under 10% in most cases, as shown in Figure 2.2. [14, 17]

2.1.1 Gasoline refining

Most refineries worldwide focus their activity on the production of transportation fuels. On a global scale, diesel comes first in terms of demand followed by gasoline [19]. Europe shows the same pattern, 40% of the crude is refined for diesel production while only 18% of the oil serves for gasoline synthesis [20]. Conversely, the US has historically preferred gasoline powered vehicles rather than diesel cars. This reflects on their refineries' output, with roughly 45% of the final product being gasoline [21].

The typical configuration of a refinery for gasoline production is shown in Figure 2.3.

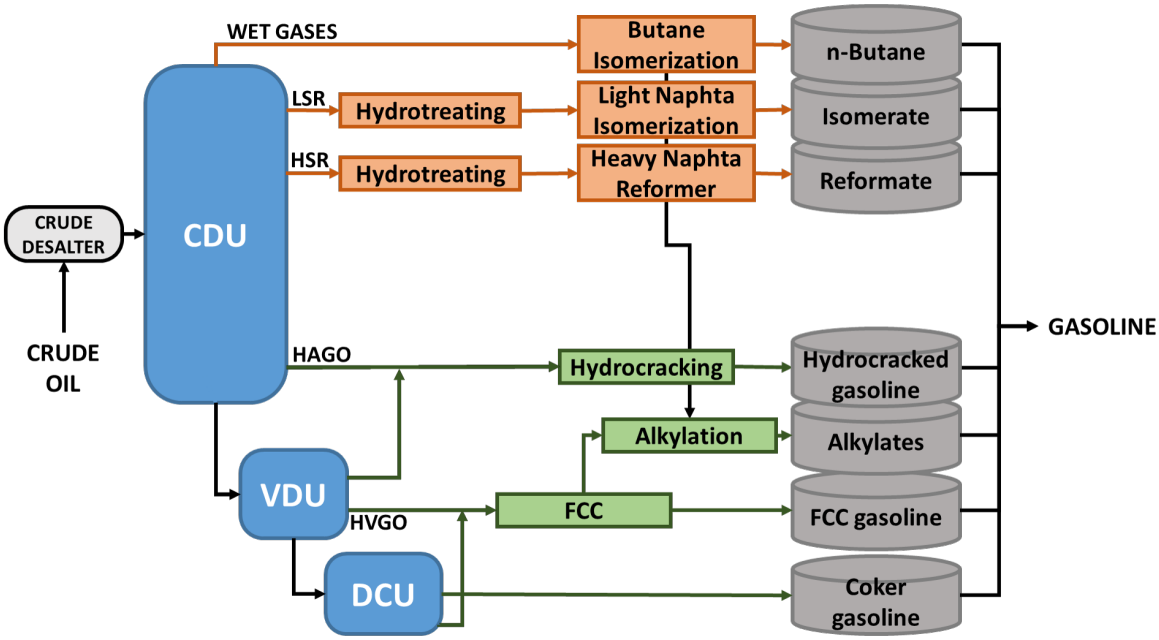

Figure 2.3 Simplified refinery layout for gasoline production

After desalination, the crude oil enters the crude or atmospheric distillation unit (CDU). The CDU separates the crude oil into fractions or cuts according to their boiling ranges, light fractions moving to the upper part of the column and heavier cuts exiting at the bottom. Part of the fractions are used for the formulation of a single product while others are further split into several streams for different purposes in the refinery. Next, the already divided fractions are sent to downstream processes. In the case of gasoline, most of these operations aim to improve fuel quality by raising octane number and removing impurities, which confers to each blending stream its own distinctive properties. Gasoline components that are common to most refineries are n-butane, isomerate, reformate, hydrocracked gasoline, alkylate, fluidized catalytic cracking (FCC) gasoline and coker gasoline. [22]

Interesting light fractions for gasoline production coming out of the CDU are shown in orange in Figure 2.3. These streams are light straight-run (LSR) naphta, heavy straight-run (HSR) naphta and part of the wet gases. Straight-run naphtas undergo hydrotreating processes to remove impurities from the streams such as sulfur, nitrogen, metallic salts or saturated aromatics. LSR is then fed into the isomerization unit which

saturates benzene and promotes n-paraffin branching and conversion into iso-paraffins. HSR heads to the reformer instead, where it is converted mainly to high-octane aromatics. Butane from the wet gases undergoes isomerization and is divided in n-butane directly used in the final gasoline, and iso-butane fed to the alkylation unit. In addition to the light streams, the heavy atmospheric gas oil (HAGO) can be also treated downstream to maximize the gasoline production at the refinery. This is achieved through hydrocraking, as shown in green in Figure 2.3. [22]

The vacuum distillation unit (VDU) is located downstream of the CDU for further processing of the heavy residual fraction, as shown in the lower half of Figure 2.3. The vacuum atmosphere allows for lower operating temperatures and prevents thermal cracking. The resulting bottom fraction known as heavy vacuum gas oil (HVGO) is processed in an FCC unit to obtain lighter hydrocarbon molecules. The residual stream from the VDU enters the delayed coker unit (DCU) for thermal cracking. Part of its output can be directly blended into the final gasoline product while another fraction is sent to the FCC unit for cracking. [22]

In addition to these streams, gasoline properties are enhanced by mixing other components, typically oxygenates such as alcohols and ethers. Methyl tert-butyl ether (MTBE) is broadly used as a gasoline booster. However, due to soil and groundwater contamination issues, it is being gradually replaced by other ethers, mainly ethyl tert-butyl ether (ETBE) [23], and ethanol, which presents the advantage that it can be obtained from renewable sources [17]. The presence of oxygen atoms improves combustion behavior and reduces the emission of harmful substances such as unburnt hydrocarbons [24, 25]. Increasing the oxygen content of gasoline is a common practice during winter season, as it helps keeping carbon monoxide emissions within the limits [26]. Moreover, some oxygenates can act as octane boosters and reduce the knocking tendency of the fuel [25]. The main disadvantage of oxygenated compounds is their lower energy content which theoretically results in higher specific fuel consumption for the engines. However, it has been suggested that blending small percentages of ethanol in gasoline can enhance thermal efficiency of the engines and counterbalance lower heating values [27].

2.1.2 Gasoline as a fuel

Gasoline is typically used as fuel for SI engines. SI engines are internal combustion engines where the combustion of the air-fuel mixture is triggered by a spark. Although various configurations are available, such as rotary engines, reciprocating engines are the most popular design, especially for automotive applications. Reciprocating engines, also known as piston-and-cylinder engines, rely on pressure fluctuations derived from the combustion of the fuel that allow the back-and-forth movement of the piston inside the cylinder. The linear motion is then transformed into rotation through the connecting rod and the crankshaft. [15]

The gasoline combustion cycle or Otto cycle is divided in four different stages, namely intake, compression, power and exhaust. Two-stroke engines perform two of these stages simultaneously and complete one cycle every revolution. On the other hand, four-stroke

engines perform each stage separately and require two rotations of the crankshaft to complete one combustion cycle as shown in Figure 2.4. Two-stroke engines find application in outdoors tools, such as chainsaws and lawn mowers, as outboard engines or in small motorcycles. Advantages of this type of engines are low cost, simplicity, robustness and high specific power. However, for larger applications, drawbacks tend to outweigh the benefits as losses become too important. Thermal efficiency and specific fuel consumption cannot compete with those of the four-stroke engine. Furthermore, both geometry and oil system play against emission levels in two-stroke engines. [15]

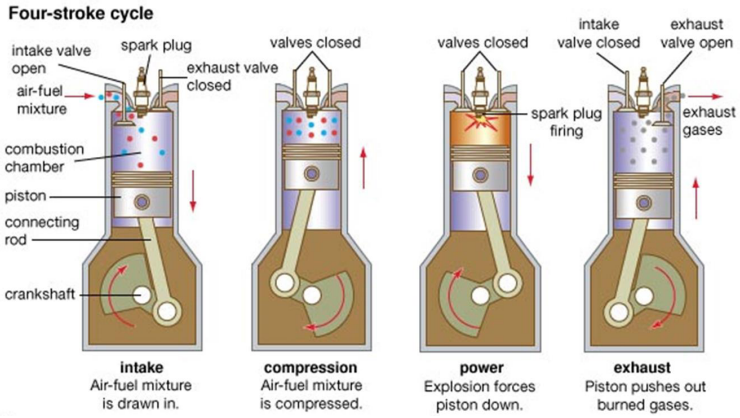

Figure 2.4 Otto cycle in a four-stroke engine [28]

The Otto cycle starts with the intake stroke. When the intake valve opens, fresh air-fuel mixture — or simply air in the case of direct injection engines — enters the combustion chamber as the piston moves downwards. This admission phase is controlled by the pressure difference between the intake port and the combustion chamber. During the second stroke, the compression stroke, both intake and exhaust valves remain closed. The piston moves now from bottom dead center (BDC) to top dead center (TDC) compressing the fuel-air mix and rapidly increasing its temperature. In an ideal cycle, combustion is an immediate process, however, in a real cycle ignition timing is usually advanced so the heat release occurs with the piston as close to TDC as possible in order to maximize power output. After the spark, the air-fuel mixture expands pushing the piston downwards in the so-called power stroke, and slightly before reaching BDC for the second time, the exhaust valve opens. The exhaust gases from the combustion exit the chamber pushed by the piston, which is now moving upwards and completing the exhaust stroke. It is a common practice to keep the exhaust valve open after reaching TDC, simultaneously with the intake valve for the proper intake of the fresh mixture and to favor the cleaning of the combustion chamber from combustion residues. [15]

2.1.3 Gasoline properties

In order to ensure the smooth performance of a SI engine, the fuel must meet certain requirements. In Europe, commercial gasoline must comply with the European directive 2009/30/EC [29] and fuel quality specifications are covered in standard EN228 [18], presented in Table 2.1. Each country may as well publish its own additional legislation, such as regulation 1206/2010 in the case of Finland [30].

Table 2.1 Main properties included in the European standard EN 228 for gasoline fuel quality [18]

	Gasoline 95		**Gasoline 98**	
	Minimum	Maximum	Minimum	Maximum
Chemical composition				
Aromatics [% v/v]	-	35.0	-	35.0
Olefins [% v/v]	-	18.0	-	18.0
Benzene [% v/v]	-	1.0	-	1.0
Oxygen [% m/m]	-	3.7	-	2.7
Ethanol [% v/v]	-	10.0	-	5.0
Ethers C_{4+} [% v/v]	-	22.0	-	15.0
Density				
Density at 15°C [kg/m^3]	720	775	720	750
Volatility				
Vapor pressure summer [kPa]	45	70	45	70
Vapor pressure winter [kPa]	60	90	60	90
Vapor lock index	-	1250	-	1250
Distillation				
Evap. at 70°C (summer) [% v/v]	22	50	20	48
Evap. at 70°C (winter) [% v/v]	24	52	22	50
Evap. at 100° [% v/v]	46	72	46	71
Evap. at 150°C [% v/v]	75	-	75	-
Final boiling point [°C]	-	210	-	210
Residue [% v/v]	-	2	-	2
Auto-ignition				
RON	-	95	-	98
MON	-	85	-	87
Impurities and metal content				
Lead content [mg/L]	-	5.0	-	5.0
Sulphur content [mg/kg]	-	10.0	-	10.0
Other properties				
Oxidation stability [min]	360	-	360	-
Gum content [mg/100mL]	-	3	-	4
Copper strip corrosion	-	1	-	1

Chemical composition

As previously illustrated by Figure 2.1, commercial gasoline is a mixture of hundreds of different compounds. Detailed regulation regarding chemical composition does not exist, but limitations apply exclusively to certain components that compromise the lifespan of the engines or pose a danger for human health or the environment. Some examples are limitations regarding oxygen content, ethanol, aromatics, olefins or benzene, as listed in Table 2.1. Other aspects related to chemical composition are not regulated but they should be taken into consideration. Carbon to hydrogen ratio for instance, determines to a great extent the combustion characteristics of the fuel, such as fuel-to-air ratio or adiabatic flame temperature. Overall, the chemical composition of the gasoline is the ultimate responsible for the rest of the properties of the fuel. [15]

Density

Density measures the mass-to-volume ratio of a substance and in the case of gasoline it is typically reported at 15.6°C. Density is connected to the volumetric energy content

of a fuel and alternatively, it can be expressed in terms of relative density or specific gravity. Dense fuels contain more energy per volume unit, therefore favoring fuel tank size reduction. Moreover, gasoline density must be controlled for proper functioning of pumps and injectors if they exist, as well as correct atomization of the fuel and consequently smooth combustion. Premium gasolines with higher octane rating generally show higher densities values due to the presence of greater aromatics fractions. [31]

Viscosity

Viscosity quantifies the resistance of a fluid to deformation. Similar to density, it conditions the performance of different engine components, having major impact on the injectors. If the viscosity is excessive, injectors are not capable of atomizing the fuel properly and large drops are delivered to the cylinders. The result is poor combustion with increased emissions and specific consumption. In the opposite case, low viscosity fuels do not provide enough lubrication for pumps and injectors. This results in damaged components but also leads to leakages.

Volatility

Gasoline volatility properties must assure satisfactory performance of spark-ignition engines under normal service conditions in terms of start-up and warm-up, acceleration and throttle behavior. High volatility may lead to vapor lock and other severe issues. Fuel vaporization in pumps, ducts or injectors reduces fuel flow to the cylinders leading to undesired starting and operation of the engine. By contrast, low volatility causes hard start and uneven distribution of the fuel among the different cylinders in the case of old vehicles with carburetor, or within each cylinder in the case of direct injection. Furthermore, volatility requirements depend on atmospheric conditions and acceptable ranges must be adjusted seasonally to guarantee optimal performance of the engine all year around. Volatility of gasoline is frequently measured by the Reid vapor pressure (RVP) and distillation curves. [31]

RVP is defined as the absolute pressure exerted by a vapor in thermodynamic equilibrium with the liquid phase at 37.8°C (100°F) and vapor-to-liquid ratio of 4:1. RVP differs from true vapor pressure (TVP) as it takes into account any dissolved air and moisture in the vapor phase. Several methods have been developed to obtain experimental values of the RVP and they can be classified in two groups. The first type employs time-consuming phase equilibrium calculations, while the second type requires full distillation curves which are not always available. In any case, obtaining RVP results a tedious process, hence the attempts to develop simple correlations that use readily available properties such as specific gravity or boiling points [31, 32].

Distillations curves pair temperatures and volumetric evaporated fractions at such temperatures. They are typically obtained using gas chromatography analysis or distillation test methods [15, 31]. For European gasoline products, evaporated volumes at 70°C, 100°C and 150°C (E_{70}, E_{100}, E_{150}) and final boiling point are representative and useful parameters for blending purposes. E_{70} is a reflection of the front end of the distillation curve connected to engine startup, risk of vapor locking and evaporative emissions [33]. E_{100} is a measurement of mid-range volatility that influences Cold

Weather Driveability (CWD) and gives information regarding acceleration behavior of a hot engine under a load [34]. E_{150} and the top range of the distillation curve measure tendency of combustion deposits formation and oil dilution and serve for fuel economy optimization in the hot engine [33].

Auto-ignition characteristics

Auto-ignition behavior is perhaps the most important quality indicator for gasoline and it can be expressed using different fuel properties. When it comes to SI engines, suitable fuels are those with poor auto-ignition characteristics which prevent knocking phenomenon from happening. Ideally, combustion should start at the spark plug and propagate steadily until all the fuel has been consumed. Conversely, engine knocking consists on the ignition of the fuel ahead from the flame front. This abnormal combustion can be avoided by choosing fuels with high octane rating, an indicator of the resistance to auto-ignition, in addition to the right geometry for both the chamber and the piston and correct spark timing. [15]

The octane number of a fuel must be determined in a Cooperative Fuel Research (CFR) test engine, a single cylinder four-stroke gasoline engine with variable compression ratio. Primary Reference Fuels (PRF) with known octane numbers are used as a reference. Iso-octane, with very high resistance to auto-ignition, is assigned an octane number of 100 and n-heptane, prone to knocking, has an octane number of 0 by definition. The octane number of any blend of these two fuels behaves linearly on a volumetric basis. Initially, the test was carried out in mild conditions for the engine, at 600 rpm and inlet air temperature fixed at 52°C. The result of this test receives the name of Research Octane Number (RON). Later on, Motor Octane Number (MON) was introduced trying to replicate more accurately the conditions of the engines in the real world. To obtain the MON, fuels are tested at higher speeds, 900 rpm, and higher temperatures, 149°C downstream of the carburetor. Although PRFs show the same RON and MON, complex mixtures and real gasolines behave differently under different testing conditions. This introduces the concept of octane sensitivity (S), which measures the difference between RON and MON. [31, 35]

$$S = RON - MON$$

The Antiknock Index (AKI) is calculated as the arithmetic mean of RON and MON. It is the common way to measure antiknock quality for commercial gasolines sold in the United States since it is considered to reflect more accurately the performance of a fuel in real situations. [31]

$$AKI = \frac{RON + MON}{2}$$

The reason why different fuels present different octane sensitivity values is still quite unclear. Leppard [36] related low sensitivity to the existence of a two-stage ignition process. Most paraffinic fuels experience low-temperature heat release (LTHR) followed

by negative-temperature coefficient (NTC) behavior and high-temperature heat release process. On the other hand, aromatics and olefins do not exhibit two distinct phases during ignition and are classified as high sensitivity fuels. Kinetic models [37,38] showed different pressure-temperature curves for RON and MON testing methods. Lower temperature in RON trajectories lead to stronger LTHR events, while MON's higher temperatures prevent the fuel from entering that region.

Octane index (OI) was introduced by Kalghatgi [39] together with a new experimental parameter K that evaluates engine operating conditions and correlates RON and MON. OI is a more representative octane measure for modern engines, with intake temperatures quite below those in the standard tests.

$$OI = (1 - K) \cdot RON + K \cdot MON = RON - K \cdot S$$

Ghosh et al. [40] identified 57 molecular lumps that can potentially describe the composition of any gasoline stream. Some lumps correspond to single molecules, while others refer to a group of compounds with similar octane numbers and blending behavior. Table 2.2 gathers all 57 lumps along with the RON and MON, either of the neat compound, or averaged for all the molecules under each category.

Table 2.2 MON and RON for common molecular lumps found in gasoline [40]

	RON	MON		RON	MON
Paraffins			**Naphthenes**		
n-butane	94	89.6	cyclopentane	100	84.9
isobutane	102	97.6	cyclohexane	82.5	77.2
n-pentane	62	62.6	m-cyclopentane	91.3	80
i-pentane	92	90.3	C_7 naphthenes	82	77
n-hexane	24.8	26	C8 naphthenes	55	50
C_6 monomethyls	76	73.9	C_9 naphthenes	35	30
2,2-dimethylbutane	91.8	93.4			
2,3-dimethylbutane	105.8	94.3	**Aromatics**		
n-heptane	0	0	benzene	102.7	105
C_7 monomethyls	52	52	toluene	118	103.5
C_7 dimethyls	93.76	90	C_8 aromatics	112	105
2,2,3-trimethylbutane	112.8	101.32	C_9 aromatics	110	101
n-octane	-15	-20	C_{10} aromatics	109	98
C_8 monomethyls	25	32.3	C_{11} aromatics	105	94
C_8 dimethyls	69	74.5	C_{12} aromatics	102	90
C_8 trimethyls	105	98.8			
n-nonane	-20	-20	**Olefins/Cyclic Olefins**		
C_9 monomethyls	15	22.3	n-butenes	98.7	82.1
C_9 dimethyls	50	60	n-pentenes	90	77.2
C_9 trimethyls	100	93	i-pentenes	103	82
n-decane	-30	-30	cyclopentene	93.3	69.7
C_{10} monomethyls	10	10	n-hexenes	90	80
C_{10} dimethyls	40	40	i-hexenes	100	83
C_{10} trimethyls	95	87	total C_6 cyclic olefins	95	80
n-undecane	-35	-35	total $C_7^=$	90	78
C_{11} monomethyl	5	5	total $C_8^=$	90	77
C_{11} dimethyls	35	35			
C_{11} trimethyls	90	82	**Oxygenates**		
n-dodecane	-40	-40	MTBE	115.2	97.2
C_{12} monomethyl	5	5	TAME	115	98
C_{12} dimethyls	30	30	EtOH	108	92.9
C_{12} trimethyls	85	80			

In addition to that, this study presents two important trends for octane numbers in hydrocarbons. First, octane number decreases with increasing carbon number, and second, for a given carbon number, branching has a positive impact on the resistance to auto-ignition of the fuel. These trends are presented in a more detailed way in Figure 2.5 and Figure 2.6 for paraffinic and aromatic hydrocarbons, respectively.

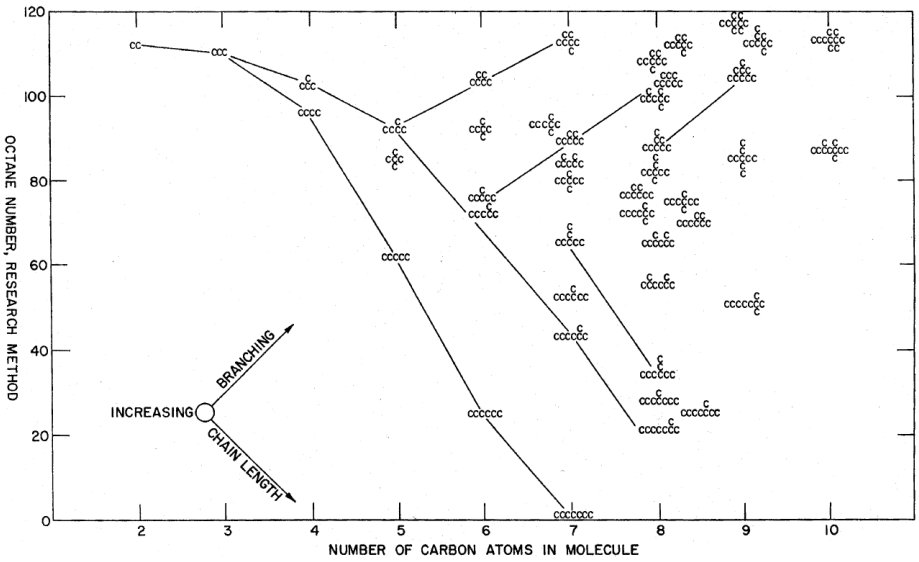

Figure 2.5 RON variation for paraffinic compounds as a function of their carbon number and branching level [40]

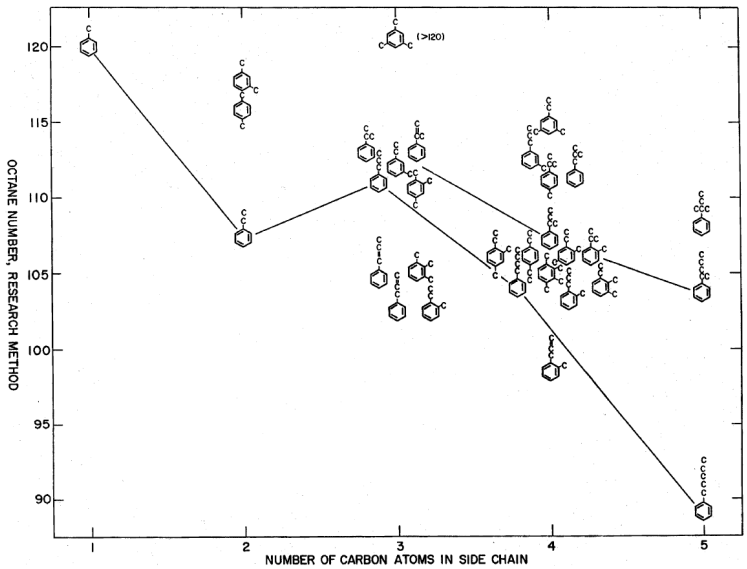

Figure 2.6 RON variation for aromatic compounds as a function of their carbon number and branching level [40]

Heating value

The popularity of gasoline for road transportation, and in general of fossil fuels, lies in its high energy density. The complex chemical composition of gasoline provides a large

amount of energy for a small volume. The average lower heating value of gasoline is between 44 MJ/kg, or 33 MJ/L on a volume basis. [15]

Impurities and metal content

Gasoline must be cleaned from impurities and pollutants for various reasons, such as avoiding engine malfunctioning, reducing environmental impact or preventing health issues in human beings.

One of those impurities is sulfur, which is naturally present in crude oil and must be removed during the refining process to avoid high concentrations in gasolines and other products. Sulfur oxides are harmful compounds for the environment and are responsible for acid rain. Moreover, sulfuric acids and other sulfur-containing species inhibit the performance of automotive three-way catalysts based on palladium and platinum [41].

Among other metals, lead poses a risk for human health and is limited to 5 mg/L [18]. Tetraethyllead (TEL) was broadly used as an octane booster during the past century, but nowadays leaded gasolines are banned in most countries [15].

2.2 Alternative liquid fuels

Increasing awareness about the impact of burning fossil fuels on the environment and more stringent emission legislation have created interests and opportunities for alternative fuels to enter the market. Biofuels have existed for decades, however they have seen an increasing popularity lately and more research effort and resources have been put on their development to reach mature commercialization stage.

Quite recently, and partly related to the increasing problem caused by overconsumption of plastics and the resulting waste streams, the idea of turning this waste into fuels has also gained attention.

2.2.1 Alternative feedstock

The use of biomass as an energy source is nothing new. The earliest humans already used wood as a form of solid biomass thousands of years ago for heating and cooking purposes. Liquid fuels came later. As early as in the 18^{th} century, whale oil and other plant derived oils were burnt to light up streets and houses. Furthermore, Rudolf Diesel designed the first diesel engine in 1897 to run on peanut oil. However, with the rise of the petroleum industry, all these products were replaced by their fossil-based equivalents, more affordable, abundant and convenient. [42, 43]

Nowadays, alternative feedstock for fuel production still refers mainly to biomass, although other types of raw materials are getting growing attention. Roughly 90% of the biomass worldwide can be classified as lignocellulosic biomass, formed by three main components, namely cellulose, hemicellulose and lignin. The remaining 10% consists on lipids and starchy material. Traditionally biofuels have been classified using

generations, but with the rise of non bio-based materials this clasification has become more flexible to accommodate new feedstock as shown in Figure 2.7. In any case, these generations attend not only to the nature of the feedstock but also to the type of conversion process that is used. It is also usual to differentiate between conventional and advanced fuels. [42]

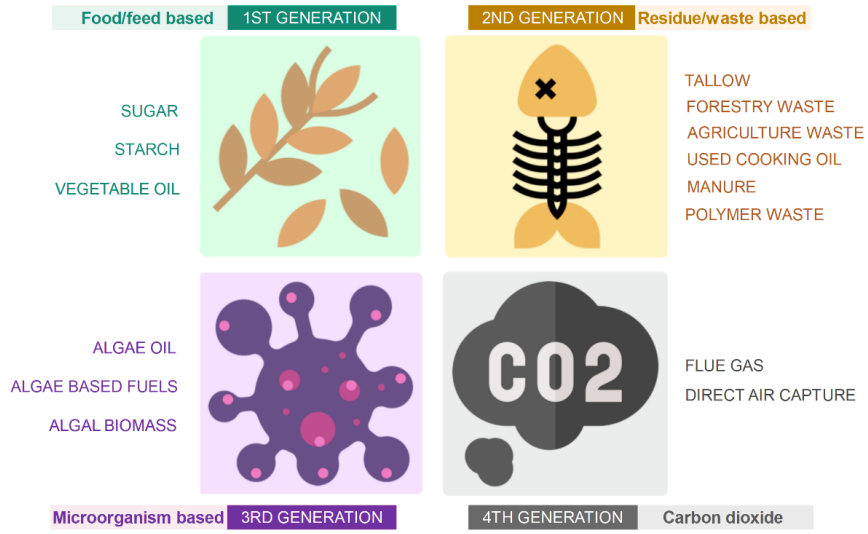

Figure 2.7 Classification of raw materials for alternative fuels by generation

First generation

First generation fuels are mainly conventional biofuels such as bioethanol, biodiesel and straight vegetable oils (SVO). Ethanol is obtained using fermentation while biodiesel is produced through transesterification, both mature and well-known processes. These fuels are obtained from edible crops that have been grown in cultivable land using conventional conversion and processing techniques. A main concern regarding first generation biofuels is land use change and competition and subsequent food price escalation. Although other generations also face the land use change challenge, the issue here is magnified by a characteristic low energy yield. [44]

First generation technologies have reached mature commercialization stage and are able to compete with fossil-based fuels. Bioethanol is mainly produced from starch and sucrose-rich feedstock, like corn and sugarcane as reflected in Figure 2.7. After a pretreatment stage, raw materials undergo hydrolysis and fermentation. The output is a mix of ethanol and by-products that must be separated and upgraded. Upgrading includes different kinds of processes, such as purification and dehydration. [45]

All gasolines in Europe typically include around 5% ethanol (E5) to improve their antiknock performance and blends with 10% ethanol (E10) are also widely available. Despite the lower energy content, their use does not largely affect performance or increase specific consumption of SI engines, since adding oxygenates can have a positive impact on fuel efficiency [27,46]. Flexible-fuel vehicles (FFV) are designed to run on any ethanol-gasoline blend, although they are typically optimized for 85% ethanol content mixtures (E85) [47].

Second generation

Second generation fuels theoretically avoid the "food versus fuel" dilemma by exploiting non-edible raw materials. Together with the third and fourth generations, they conform the so-called advanced biofuels. In addition to the use of alternative raw materials, the word "advanced" also refers to more sophisticated conversion routes. Fuels obtained from second feedstock started to be produced at full commercial scale in 2015 and currently production sites can be found in most regions worldwide [48].

This group includes both energy crops and different types of waste as raw materials. Energy crops are vegetable species with optimal characteristics for fuel production that can grow in marginal land [49]. Ideally, these crops show fast growth rates and high energy densities, together with low water consumption and inexpensive processing. In some cases, genetic modification is used to enhance yield and reduce costs, hence increasing overall profitability. Nevertheless, the use of energy crops does not fully cease the debate on land use change [50]. If not planned correctly, their growth can reduce the available space for food cultivation and result in increasing prices too [51]. The alternative to energy crops is using waste materials as showed in the top left corner of Figure 2.7. This includes forestry residues, animal fats, used cooking oils or waste streams from agricultural industry, but also plastics [44].

Plastics are polymeric synthetic materials, mainly used for packaging purposes. For plastics to be used as fuel raw material, first the long molecular chains must be broken down using degradation processes that involve high pressures and temperatures. Depending on the technology, solid, liquid or gaseous products are obtained in variable proportions. After upgrading, the resulting products present similar appearance, composition and properties to those of fossil fuels. Liquid fractions can be further processes to obtain gasoline-like fuel, besides other typical refining fractions, such as diesel or jet fuel. Some of the advantages of plastic derived fuel are low sulphur content, which prevents harmful emissions and acidity issues, and low water content, that reduces corrosion problems. On the other hand, plastic materials collected from municipal solid waste are related to high ash content, especially when low density polymers are involved. [52]

Third generation

Third generation fuels are based on microorganisms, including different types of micro and macro algae and bacteria as expressed in Figure 2.7. Microalgae have high photosynthetic efficiency and show faster growth rates than any other terrestrial plant species. While corn or sugarcane are annual crops, algae require just a few weeks to grow — or even less under the optimal cultivation conditions — easily resulting into more than 20 biomass cultures per year. They are a great source of lipids, some species showing up to 70% oil content on a dry basis, such as *Botryococcus braunii*, *Nannochloropsis* or *Schizochytrium sp.* High oil percentages are achieved at the expense of biomass content; hence, more balance species like some *Chlorella* strains are often preferred depending on the application. [44]

As an additional advantage, algae no longer need to be cultivated using arable land, but they can grow in water bodies or infertile soil instead. Moreover, their cultivation shows

potential to be combined with wastewater treatment lowering fresh water consumption. However, energy consumption is remarkably higher and the risk of biomass contamination is a main issue with open systems used in large-scale production sites. [44]

Fourth generation

Fourth generation fuels aim for carbon neutrality or even carbon negativity with the help of synthetic biology technology. This includes genetically modified algae species in combination with clean electricity sources, mainly solar energy, for production of fuels and chemicals. Genetic modifications enhance photosynthetic efficiency, increase light penetration and reduce photoinhibition [53]. These changes lead to higher CO_2 fixation and new pathways for fuel production.

The production of fuels from CO_2 that do not rely on any living organism is also possible. Several conversion technologies have already been suggested in literature to obtain liquid hydrocarbons based on CO_2 hydrogenation [54].

2.3 Renewable gasoline

Renewable gasoline, also known as bio-gasoline or "green" gasoline, refers to liquid fuels resembling conventional fossil-based gasoline in both composition and behavior, but instead obtained from a renewable resource. These fuels are also regulated by legislation and standards, such as Directive 2009/30/EC [29] or EN 228 [18] in Europe. Unlike bioethanol, renewable gasoline is a drop-in fuel that can be used in SI engines without performing any modification. Moreover, it can be obtained from sources that do not compete with food production.

Possible raw materials for renewable gasoline production have already been discussed in section 2.2.1 and include those pictured in Figure 2.7: lignocellulosic biomass, algae, CO_2 and different types of waste. Biological and catalytic pathways are still under development but open the possibility of lignocellulosic drop-in fuels to replace fossil gasoline. The main challenges relate to process integration, such as the need of pretreatment and post processing for complete conversion of the biomass and catalyst poisoning, and low productivity [55]. Algae can be converted in bio-oil and further processed in bio-gasoline, or combined with CO_2 capture. Regarding waste streams, gasoline production has been suggested using by-products from agricultural activities [56–58], animal fats [59, 60], sewage sludge [61]or plastics [52], among many others.

2.3.1 Conversion pathways

To transform living matter into green gasoline, raw materials must be extensively treated. Despite some of the processes being complex and requiring several steps, the advantage of this approach is that almost any feedstock can be transformed into gasoline-like fuel if the correct pathway and after-treatment steps are chosen. Two common advanced conversion technologies are pyrolysis and hydrothermal liquefaction (HTL).

Pyrolysis is the thermochemical degradation of biomass in absence of oxygen to obtain liquid, solid and gaseous products. Modifying the operational conditions affects the phase distribution. Slow pyrolysis, meaning long residence times and slow heating rates, favors the formation of solids and gases. On the other hand, fast pyrolysis, with short residence times of vapors and fast biomass heating rates, increases the yield of liquid products. [62]

During HTL, biomass is treated in pressurized water. Temperature must be kept high enough to ensure liquid or supercritical state of the water. HTL is suitable for wet biomass with high water content. It is a promising technology for algae processing into fuels, as well as for other wet feedstock as sewage sludge. The outcome of the process is known as biocrude. Although biocrude properties are closer to crude oil than those of pyrolytic oil, liquid upgrading is still required to meet the technical criteria from refining processes. HTL is often coupled with catalytic hydrotreating for this purpose. [63]

2.3.2 Properties and availability

The liquid fraction obtained from pyrolysis is known as bio-oil. As a result of high concentration of oxygenate compounds, bio-oil presents lower energy content and higher viscosity and corrosiveness than mineral oils [64]. Additionally, bio-oil from algae pyrolysis presents high content of nitrous species as well as aldehyde-ketones molecules derived from protein degradation [65].

Biocrude from HTL presents higher heating value and lower oxygen content than products from pyrolysis. Nonetheless, the properties of biocrude still differ from crude oil and vary depending on the raw material and other factors. Even after upgrading, biocrude from different feedstock present different characteristics. Jarvis et al. [61] compared biocrude obtained through HTL from pine wood, algal mass and sewage sludge. While algae and sewage sludge yield similar products, biocrude from pine wood had a different composition, possible as a result of higher lipid, protein and cellulose content in the raw material.

The variable composition combined with low or intermittent availability of alternative raw materials has been the main barrier for the deployment of these technologies and the commercialization of green hydrocarbons until now. Petroleum reservoirs contain large volumes of crude oil making it possible for refineries to secure the supply of raw material and guarantee production for long periods. In addition, low variability on the properties and composition of the crude simplify planning and scheduling. On the other hand, alternative feedstock is available at lower rates and continuous supply of biocrude requires the combination of different streams with variable composition. However, making use of appropriate post-processing techniques it is possible to deliver homogeneous biocrude streams regardless of the original material source.

3 Modeling tools

Moving away from conventional transportation fuels and instead introducing clean energy sources in existing refineries requires changes in the production schemes. Variable feedstock calls for more flexibility and rapid response to changes in the supply. This evidences the need for advanced tools and models capable of managing these fluctuations without negative implications on the quality of the final products.

The first part of this chapter presents the current production strategies followed by most refineries worldwide and how blend properties are predicted. The section focuses on properties with non-linear behavior, specially octane numbers and octane sensitivity. The second half of the chapter deepens on the core of this master's thesis and several Machine Learning (ML) algorithms are introduced for their use in fuel blend property prediction.

3.1 Blend property prediction

Planning and scheduling production is a critical task for a refinery and adopting the optimal operating strategy can save millions of euros per year. On the other hand, making the wrong decisions can result in under-graded products and the need of expensive additives to meet the market requirements.

Most refineries plan their production several months in advance. A successful planning stage should achieve a balance between inputs and outputs while providing customers with the right products and keeping in mind the profitability of the refinery. Due to the complexity of the problem, there is no single standard approach, instead each refinery tends to adopt and implement its own strategy. Nevertheless, it is an extended practice the use of mixed integer linear programming (MILP) or mixed integer non-linear programming (MINLP) solvers to maximize economic profit. In addition to continuous variables, mixed integer programming relies on variables that can only take integer values. Frequently, binary variables are used to represent whether a unit, process or other element is operating or not. Integer variables can also define thresholds, for instance, the maximum number of streams blended for a certain product. [66, 67].

These models rely on different types of data. On the one hand, they need information regarding the inputs of the refinery, such as crude oil, streams from other production sites or purchased additives, as well as inventory data. On the other hand, product specifications and demand levels must be also provided. In addition to that, it is necessary to know any type of constrains regarding product quality or operating conditions for the different units. Last, an objective function must be defined to allow the model to optimize the production for maximum economic profit. When all the objective functions and constrains are linear, MILP formulations can be used, otherwise, it becomes a MINLP problem. Overall, either linear or non-linear approaches return reasonably accurate solutions to the problem. [67]

One of the main challenges in this process is the prediction of the final properties of

the blends. As already mentioned in section 2.1, commercial products are not obtained from a single unit but they are the result of combining several streams in the right proportions. Moreover, one refinery may supply several market locations, which in the case of gasoline translates to several outputs with variable specifications that change throughout the year to meet seasonal requirements. Many linear models have been proposed to express the behavior of physical and chemical properties of petroleum products. However, these models are usually simplifications of the reality as linear interactions among chemical molecules seldom occur. [67]

Some properties can actually be predicted with linear models, for instance those related to chemical composition. Aromatics and olefins content, oxygen concentration or hydrogen-to-carbon ratio show a linear behavior in blends. In these cases, errors are mostly attributable to the uncertainty of standard procedures to calculate the property for the single components or streams. In other cases, the blending is not linear, however, a linear approach does not result into a significant error, such as for density estimations. However, there are some properties whose behavior is not only far from linear but unpredictable. Auto-ignition related properties are a representative example of this type of behavior. RON, MON and subsequently S, show different blending responses depending on the molecules involved in the mixture. Other examples are distillation curves and vapor pressure. [67]

The introduction of renewable compounds in conventional refineries may pose a challenge for traditional refining schemes and blending strategies. Exact blending behaviors are known for very few molecules and existing correlations are the result of the experimental work. Drawing attention again to octane numbers, the reasons behind non-linear interactions between molecules are quite unclear, as previously explained in section 2.1.3. Moreover, alternative bio-based and synthetic products can greatly differ in composition from fossil-based ones, in part due to higher concentrations of oxygenates and nitrous compounds [61, 65] which shape their performance in a different way. Existing models cannot fully capture the nature of these components correctly and fail in predicting final properties with accuracy.

Published studies in this field can be classified in two main groups, depending on whether the attention is put on predicting the octane number for single molecules or for gasoline blends.

For the prediction of octane numbers for pure compounds, studies mainly focus on finding correlations between molecular descriptors or indicators and octane numbers. As early as in 1931, Wheeler et al. [68] described knocking trends for straight-chain and branched olefins. More recently, Boot et al. [69] reviewed the chemistry of well-known octane boosters to extract generic rules on antiknock performance and investigated the effect of molecular structure on auto-ignition reactions separately for paraffins, olefins, aromatics and oxygenated compounds. Monroe et al. [70] reported hyperboosting effects for prenol and gasoline blends, where the octane number of the blend exceeded the octane number of the neat component. Consequently, they investigated other C_5 alcohols with a similar molecular profile but the phenomenon did not prevail. Models based on the contribution of structural groups to octane numbers have also been proposed [71], as well as algorithms built on topological indexes such as number of carbon atoms [72]

or making use of infrared absorbance spectra measurements [73].

A second group of researchers have studied the interaction between molecules to try to predict the octane number of gasoline based on its composition. Foong et al. [74] experimentally defined RON and MON blending curves for ethanol and gasoline and its surrogates. Their results highlight a synergy between ethanol and paraffinic compounds, which can be beneficial for fuel design with lower aromatic content. Ghosh at al. [40] developed a detailed composition-based model to predict RON and MON for gasoline. In their model, the contribution of each molecule to the octane number of the blend is not dependent on the octane number of the neat compound but on its blending value. They found out that the blending value of a molecule varies almost linearly with the octane number of the fuel or blend it is part of. Additionally, infrared spectroscopy [75, 76], nuclear magnetic resonance [77] or gas chromatography techniques [78, 79] have also been correlated with gasoline octane numbers.

Throughout this thesis some of the most common and representative molecules in gasoline fractions are presented, as well as the interaction among them in order to predict RON, MON and S. The predictions are carried out using different models and algorithms, which also attempt to help clarifying the complex mechanisms that rule molecular interactions.

3.2 Machine Learning for model development

Mathematical representation of biological and chemical phenomena has traditionally been achieved through conventional statistics. However, the increasing availability of cheap computing power has recently opened the door for ML techniques as well. A typical definition of ML based on Arthur L. Samuel's publication [80] is "the field of study that gives computers the ability to learn without being explicitly programmed". Regarding the question this thesis tries to solve, this means a computer is able to predict the properties of a blend without explicitly inputting all the physical and chemical interactions between its components.

There is a lot of shared knowledge between ML and statistical analysis, which sometimes makes it hard to draw the line between the two fields. In fact, the area of ML borrows many concepts from statistics, as well as from other disciplines, such as information theory. Nevertheless, ML is a subfield of computer science and artificial intelligence, while statistics is a subfield of mathematics. By definition, statistical methods focus on inference and fitting problems while ML specializes on generalization and prediction of unforeseen events. However, many examples can be found where statistics are used to make projections and ML succeeds in finding the explanation behind a natural phenomenon. As a result, the same algorithm can be used in ML and statistics but with a different approach and for a different purpose. [81]

The distinction between the two fields is also reflected by the type of data they use. Statistics produce satisfactory results with small sets of "long data", that is, a greater number of observations than variables. On the other hand, ML can deal with larger

volumes of data in a wide format, meaning more input variables than subjects. But with the increasing complexity of the data, ML brings along some risks. In fact, lack or low-level interpretability afforded by most algorithms make it difficult to prove and understand relationships within the data. [81]

Depending on whether ML systems rely on human supervision during the training stage or not, they can be classified as supervised, non-supervised, semi-supervised or reinforcement learning:

- In supervised learning, the data used to train the algorithm includes the desired outputs or labels. The labels in the case of the blending problem are the values of the target properties for the blends. Typical supervised learning tasks are classification and regression. [82]
- Unsupervised algorithms are not provided with desired outputs, but they are used to find relations within the data. Main tasks covered by unsupervised leaning are clustering, association, visualization, dimensionality reduction and anomaly detection. [82]
- Semi-supervised learning algorithms combine features of both supervised and unsupervised learning. Generally, part of the data is labeled while the rest is not. A common approach is to train the algorithms in an unsupervised way and fine-tune them taking advantage of the labeled data. [82]
- Reinforcement learning is based on a reward system. The learning agent takes actions in its environment and receives positive or negative rewards for them. The agent will learn from its actions as will try to maximize the cumulative reward after each decision. [82]

Although the applied part of the thesis focuses exclusively on supervised learning and regression tasks, this chapter will also refer to classification tasks and include a brief overview of some common unsupervised learning methods. A comprehensive list of the algorithms covered in the literature review is included in Table 3.1.

Table 3.1 Classification of ML algorithms covered in this literature review

Supervised learning	**Unsupervised learning**
Linear Regression	K-means
Nearest Neighbors	Hierarchical Cluster Analysis
Support Vector Machines	DBSCAN
Decision Trees	Principal Component Analysis
Random Forest	
Artificial Neural Networks	

3.2.1 Linear Regression

Linear regression is a clear example of the strong relation between ML and statistics. Originally a statistical tool, ML has also adopted it as a prediction tool. A large number

of estimation methods are available to perform linear regression, however, for the scope of this master's thesis, only a few of them will be covered. Probably, the most used one is Ordinary Least Squares (OLS), where a linear function is obtained minimizing the squared difference between the observed and predicted values, as shown in Figure 3.1.

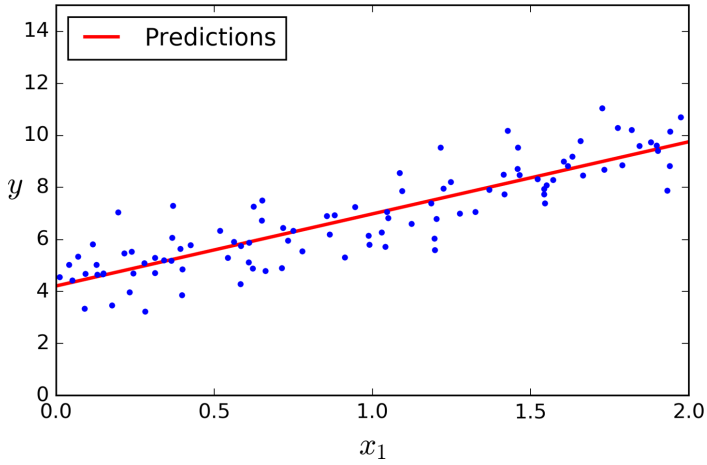

Figure 3.1 Linear Regression model predictions for a given data set [82]

When multi-collinearity and overfitting are likely to arise, Ridge and Lasso regression are often used, especially if the goal is prediction rather than inference. These two algorithms use regularization techniques that reduce variability by introducing penalty terms in the cost function. In the case of Lasso regression, L1 regularization is used and the penalty term is proportional to the value of the coefficients. This means the algorithm will try to shrink the coefficients to minimize the error. Ridge regression adopts L2 regularization so the penalty term is proportional to the square of the coefficients. This can lead to zero coefficients and help in feature selection. [83]

3.2.2 Nearest Neighbors

Nearest Neighbors algorithms are among the simplest estimators in the ML toolbox. The basic idea behind these models is that similar objects exist in close proximity or "are neighbors", as shown in Figure 3.2. In the k-Nearest Neighbors (k-NN) algorithm the number of neighbors is set by tuning the hyperparameter k, while for the radius-based Nearest Neighbors (r-NN) learning method, objects within a given radius r to the query point are taken into consideration. [84]

Unlike other supervised learning algorithms, Nearest Neighbors algorithms return non-parametric models and do not have explicit training stages. For every prediction, the model sorts all the examples in the training data set according to the distance to the query point and picks the closest ones to compute the solution. In general, the Euclidean or straight-line distance is used for the sorting, although certain problems might respond better to alternative approaches. [84, 85]

For regression tasks, the algorithm returns the average of the n closest objects, while for classification it returns the most common class among those n neighbors. This

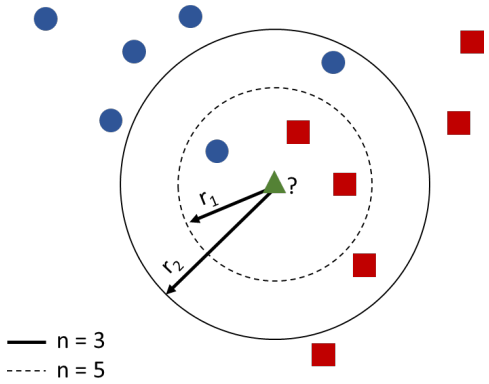

Figure 3.2 Nearest Neighbors approach to classification tasks with two classes, red squares and blue circles

process must be repeated for every new prediction, which makes the algorithm become slower with increasing volumes of data. The number of neighbors, as well as other hyperparameters, can be optimized through an iterative process, modifying the value of k or r and comparing the error on the validation set. In classification tasks, odd numbers are preferred to avoid tie situations. In general, decreasing the number of neighbors to one produces unstable models, while too large values increase the error of the predictions. [84]

3.2.3 Support Vector Machines

Support Vector Machines (SVM) are versatile and powerful machine learning resources with great popularity. SVM can be used for both classification and regression tasks, as well as a tool for anomaly detection. Furthermore, they show reasonably accurate results for complex problems even when a small or medium size data set is available. [82]

SVM classifiers define hyperplanes that separate labelled data into different classes. The algorithm is trained to maximize the minimum distance, called margin, between classes. Figure 3.3 shows data belonging to two different classes, red squares and blue circles, and possible ways to classify them.

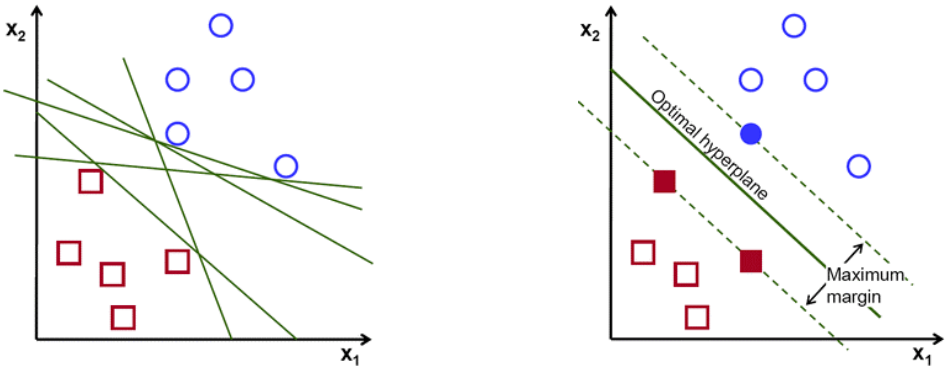

Figure 3.3 Possible hyperplanes for classification of samples (left) versus SVM approach to the same problem with margin maximization (right) [86]

Figure 3.3 compares multiple lines (2D hyperplanes) able to classify the samples correctly (left) against the hyperplane that maximizes the distance between the two classes (right). The latter is the approach followed by SVM algorithms, able to reduce the risk of misclassifying unseen data when utilizing the model in a later stage. The samples that define the maximum margin (solid markers in Figure 3.3) are called support vectors. SVM algorithms are mainly built on information provided by these support vectors, hence reducing overfitting risk. Unlike linear regression, outliers have limited or no effect on the resulting models. [87]

Similar principles apply for Support Vector Regression (SVR). In this case, the algorithm tries to fit as many instances as possible within a given margin epsilon (ϵ), as in Figure 3.4. [82, 87]

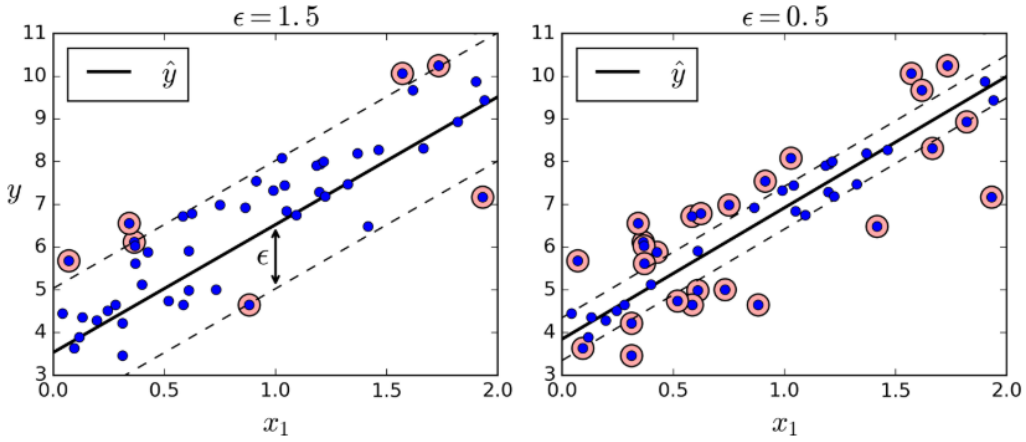

Figure 3.4 Regression analysis using an SVR algorithm and different values of ϵ [82]

To tackle non-linear problems, kernelized models can be used. Kernels transform the original input using mathematical functions and map the data into a higher dimensional feature space, as depicted in Figure 3.5. If the right kernel is chosen, the data becomes linearly separable after the transformation and a decision boundary can be fit to separate classes or perform regression analysis. The most popular kernel is the Gaussian radial basis function (RBF) kernel. [87]

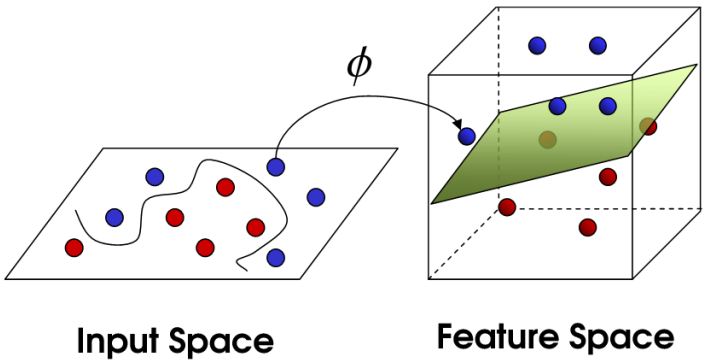

Figure 3.5 Data projection onto a higher dimensional feature space using the kernel trick [88]

Liu et al. [89] predicted the octane number for single molecules present in gasoline using Quantitative Structure Property Relationship (QSPR) models. In their work, four different techniques were compared, namely Partial Least Squares (PLS), Artificial Neural Networks (ANN), SVMs and Random Forest (RF). Best results were reported using SVM, although RF predictions were also measured as satisfactory.

Their experimental data comprised RON and MON measurements for 279 and 273 molecules, respectively, including hydrocarbons, oxygenates and nitrous compounds with octane numbers ranging from slightly negative values to over 120. Data was split into training (70%), validation (20%) and test (10%) sets and each molecule was initially represented using E-dragon software by 1667 molecular descriptors. In the following steps, the number of descriptors was greatly reduced using filter and wrapping methods. PLS was notably inferior to the other three models due to highly non-linear relations between molecular descriptor and octane numbers. ANN incurred in higher overfitting. Between SVM and RF, although the former performed better for the available data, the authors do not give a definitive explanation on which one is preferred. Nonetheless, the capacity of SVM to minimize the impact of outliers is highlighted in the text, despite the time consuming task of parameter selection.

3.2.4 Decision Trees

Decision Trees (DT) are decision support tools that use tree-like structures which are "grown" based on the probability of events to happen. The tree starts to grow from a root node, which can be identified at the top in Figure 3.6. At each internal node one feature is evaluated, with the branches coming from that leaf representing the different values that feature can take, either as a single value for discrete data or as a range for continuous features. End nodes are known as leaf nodes and they contain all the possible outputs of the model. [84]

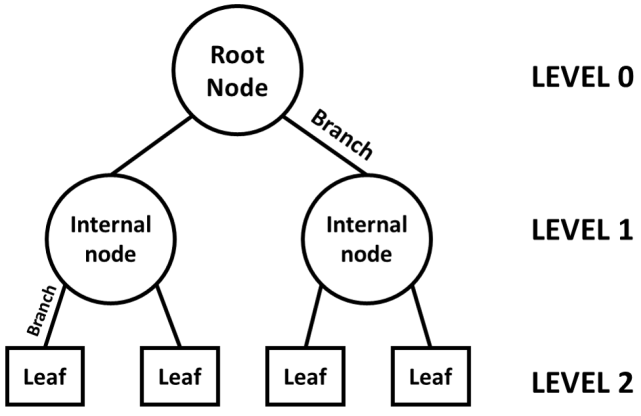

Figure 3.6 Main elements in a Decision Tree

Increasing the depth of a tree adds complexity and tends to lead to overfitting, although it might be beneficial for problems with large sets of features. In these cases, pruning can be applied to remove nodes or tree sections that have a minimum contribution to the predictive power of the tree. [84]

Decision Trees show the advantage that little data preparation is required; there is no need for data scaling or centering. Moreover, they allow for model interpretation, but simplicity is key for quality results and maximum depth should be limited. On the other hand, they are non-parametric methods, which might pose a limitation depending on the final application. [82]

3.2.5 Random Forest

Decision Trees are the fundamental component of a more complex algorithm called Random Forest. RF builds a large number of independent trees and averages the predictions of all of them as shown in Figure 3.7. In general, the increase in complexity is offset by a higher accuracy in the predictions.

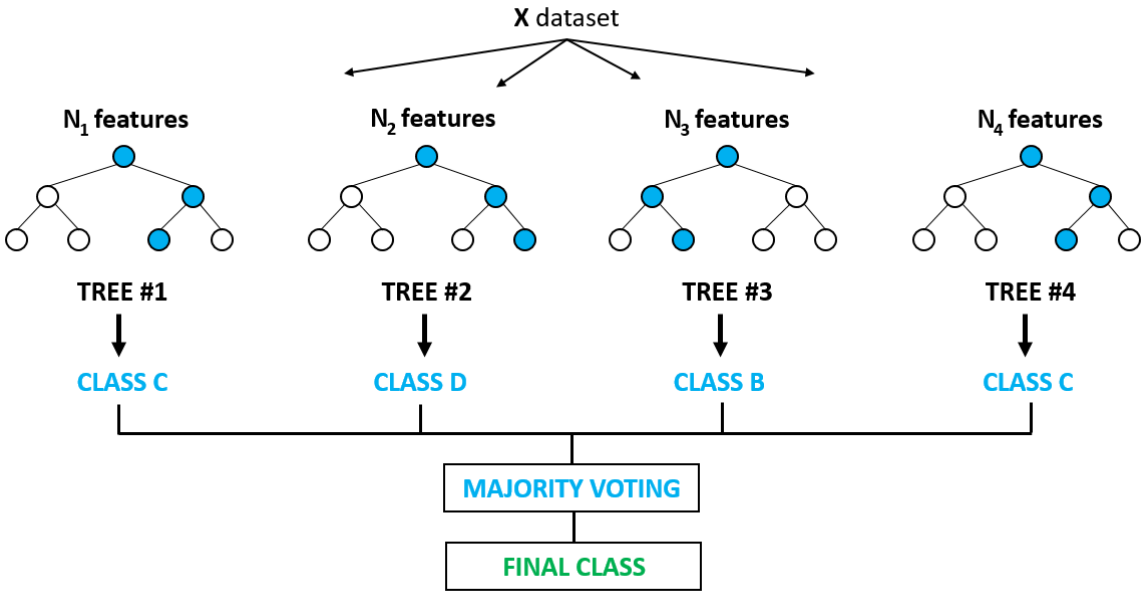

Figure 3.7 Random Forest algorithm built as an ensemble of Decision Trees [90]

Each tree in a RF should be different, yet able to make acceptable predictions. To achieve this, each tree is trained on a bootstrap sample of the initial set, where some points have been replace by duplicates of the remaining ones. Moreover, a constrain is set to the number of features each internal node can "see" and use to make the split. [84]

Lee at al. [91] used near-infrared analysis results of 379 gasoline and naphtha samples to train a RF algorithm for RON prediction. To build each tree, 8 out of 12 available spectra were selected randomly; afterwards, one third of the wavelengths in those spectra were chosen. Minimum error rates were achieved with 2000 trees. The performance of the RF model was compared to a regression by PLS showing improved accuracy.

3.2.6 Artificial Neural Networks

The brain's structure served as inspiration to create the first Artificial Neural Networks in the 1940's. Since then, they have been further developed to the point that resemblance to the biological neural systems has almost disappeared.

A biological neuron, pictured in Figure 3.8, presents three main components: the cell body, the dendrites and the axon. Neurons receive information in the form of electrical impulses through the dendrites and transmit it to the body of the cell. The information is processed and converted into an output signal that travels through the axon towards other neurons. The next neuron must receive enough strong signals to trigger its own signal and continue the process. The behavior of a single neuron is rather simple, but complexity increases when billions of them perform collectively, and the same applies to artificial neurons. [82]

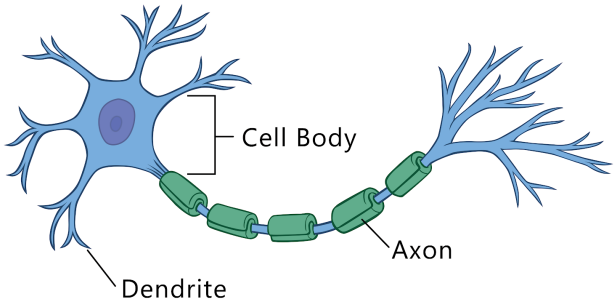

Figure 3.8 Main parts in a biological neuron [92]

One of the simplest approaches to artificial neural networks is known as perceptron, and the mechanism behind it is similar to the one explained for the human nervous system. Perceptrons are aggregations of simpler units called threshold logic units (TLUs), shown in Figure 3.9.

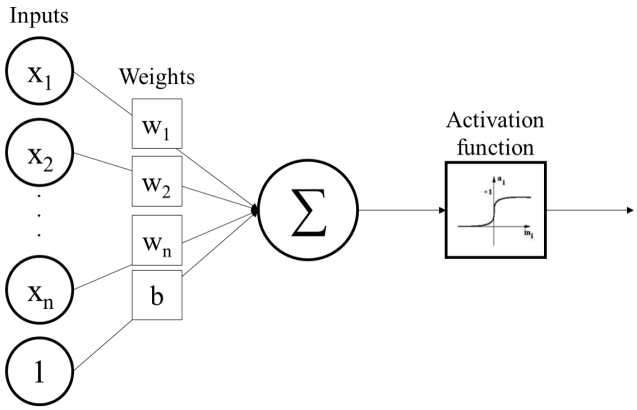

Figure 3.9 Constitutive elements of a TLU

In a TLU, the input signals correspond to numerical data fed by the user (x_1, x_2 ... x_n), each of them associated with a weight (w_1, w_2 ... w_n). In addition, a bias is often

included (b). The processing and conversion of the signals happens in the body of the unit, which computes the weighted sum of the inputs plus the bias and applies an activation function to obtain an output. Popular activation functions are step function, logistic function, hyperbolic tangent and ReLU. Perceptrons suffer from some weaknesess and are unable to solve some trivial classification problems; however, the combination of several of them in the form of a Multi-Layer Perceptron (MLP) can overcome those limitations. [82]

Typical ways of training an ANN include back-propagation. Before exposing the network to any data, the weights and the bias are initialized with some random values. Then the features of the first instance or sample are fed as input data and the network makes its first prediction sending information forward through the different layers. After that, it compares the computed value with the desired output and uses the error to update the weights and reinforce the connections through back-propagation. [82]

Albahri [93] designed several neural network structures to predict various properties for gasoline and petroleum fractions, including octane numbers. The basic input data for all models consisted on distillation curve points, however, due to the complexity of some properties, other network architectures were tested using additional information. For the RON model, predictions' accuracy improved significantly when RVP and SOA composition (saturates, olefins, aromatics) were included. Best results were obtained with a single layer feed forward neural network with 11 neurons in the input layer and 7 neurons in the hidden layer. From the 178 available gasoline samples, 85% of them were used for the training stage and 15% for testing. The test set results showed 5.4% average error and 14.9% maximum deviation. The same architecture was used to predict MON and the average and maximum error this time were 3.3% and 8.8%, respectively. The use of a single apparatus, like distillation curve in this study, to predict several properties can result of interest for modern refineries as it can save time and reduce costs.

Abdul Jameel et al. [77] developed a neural network based on nine different parameters to predict the octane number of pure hydrocarbons, ethanol blends and gasolines. Seven of those parameters contained information regarding functional groups obtained through Nuclear Magnetic Resonance Spectroscopy, while the other two were molecular weight and index branching. With the final network architecture a regression coefficient of 0.99 was achieved, which corresponded to mean absolute errors of 1.8% for RON predictions and 1.6% for MON.

3.2.7 Clustering methods

Clustering algorithms find common features among unlabeled data and group samples on that basis. Algorithms can be classified according to their clustering approach and suitability to particular data distribution types, as exemplified in Figure 3.10. Some of the most common clustering algorithms are described hereunder.

K-means algorithm uses a partition approach. It separates data in groups with equal variance and minimizes the *inertia*, or within-cluster sum-of-squares criterion,

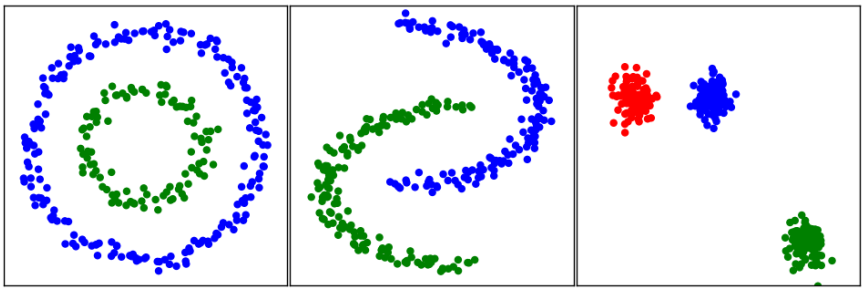

Figure 3.10 Clustering algorithm applied to different data sets [85]

to discover underlying patterns. The center of a cluster is called centroid, and the number of centroids can be adjusted through the value of parameter k. Thereafter, the algorithm finds the optimum location for each centroid through an iterative process that fulfills the inertia criterion. In Figure 3.11 k has been set to 3 and each sample has been allocated to the right centroid following the variance principle. [85]

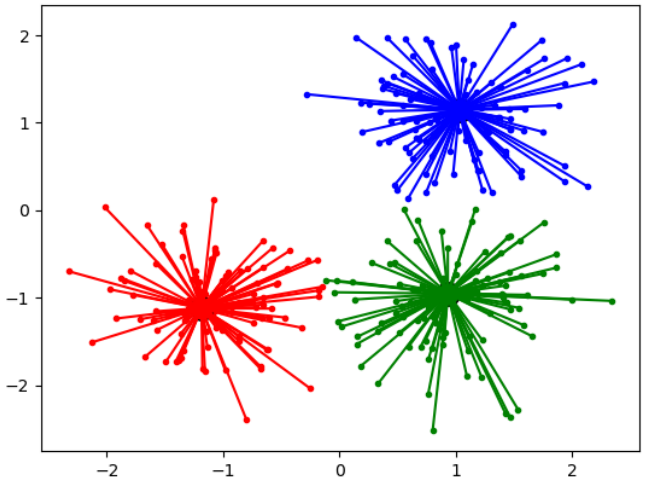

Figure 3.11 K-means is used to find three clusters within a data set minimizing the *inertia* [85]

Hierarchical Cluster Analysis (HCA) can take two different approaches. A "bottom-up" approach results into agglomerative clusters, where each observation begins in its own cluster and is paired successively with others to advance in the hierarchy. On the other hand, a "top-down" approach generates divisive clusters from a single initial group containing all the information. A common way to visualize the outcome of these algorithms is with dendrograms like the one shown in Figure 3.12, where two and three clusters are created. The y-axis gives a notion of how far apart the merged samples and clusters are. For the example in Figure 3.12 samples 1 and 4 in the green cluster share more similarities since they are closer than samples 2 and 8. [84]

Ferreiro-González et al. [94] characterized 30 commercial gasoline samples using headspace mass spectrometry and applied HCA to classify them according to their RON. In addition to HCA, they also tested Linear Discriminant Analysis (LDA) algorithm and concluded that these type of clustering methods show a high reliability level for gasoline classification tasks.

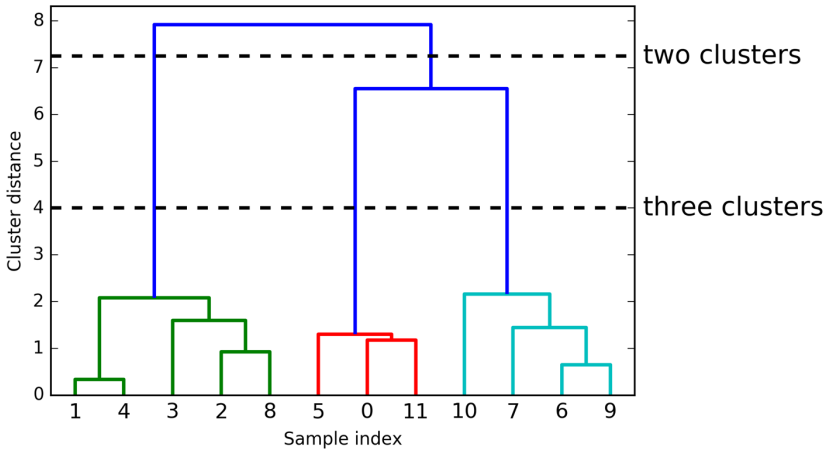

Figure 3.12 Dendrogram with dashed lines indicating splits into two and three clusters [84]

DBSCAN algorithm stands for density-based clustering method for applications with noise. To be included into a cluster, points must have a minimum number of close neighbors within a given radius as reflected in Figure 3.13 (left). For a given data set, points are classified either as *core points* (red), *reachable points* (yellow) or *noise points* (blue) following that logic. Unlike k-means or HCA, this method is less sensitive to the presence of outliers and performs better for non-convex clusters and arbitrary shapes, as those in Figure 3.13 (right). For the pictured data set, the algorithm finds five different clusters, with outliers (black markers) being left out of the classification. [84]

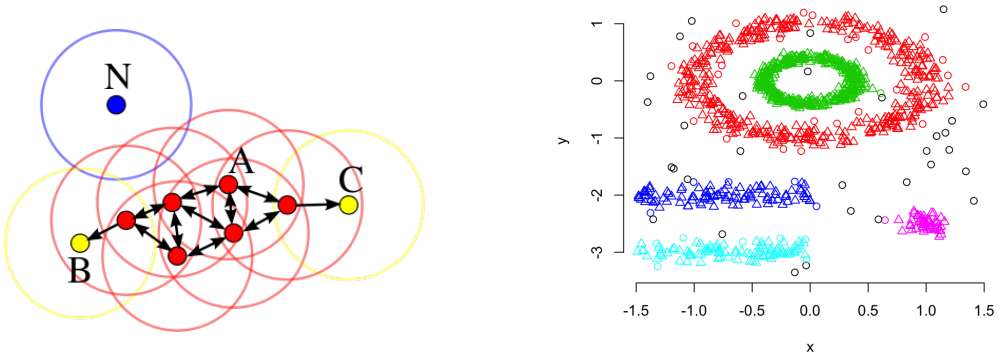

Figure 3.13 DBSCAN algorithm's working principle (left) [95] and example of performance for arbitrarily-shaped clusters with outliers (right) [96]

3.2.8 Dimensionality reduction and visualization methods

Algorithms for dimensionality reduction are useful when dealing with high-dimensional data. Reducing the number of features results into condensed data sets with less degrees of freedom where pattern recognition becomes an easier task. Moreover, these processes help with data visualization and representation and minimize computation time and loads. [82]

Dimensionality reduction can take two different approaches: feature selection and feature extraction. Feature selection algorithms look for subsets of variables within the original data set that preserve enough information to model the problem. On the other hand, feature extraction transforms the input data into completely new variables in a lower-dimensional space. [84, 97]

Liu et al. [89] followed a two-step feature selection method to develop QSPR models for RON and MON prediction of pure components commonly found in gasoline. They used a filter algorithm to remove noise in first place, followed by a wrapper to eliminate redundant features and improve model performance. The initial data set for this study consisted in 279 compounds for RON and 273 for MON, characterized using 1667 molecular descriptors. The Boruta algorithm was chosen for the filter step and it reduced the number of features in the data set to 126 and 119 for RON and MON, respectively. During the multiobjective wrapper stage, NSGA-II algorithm further reduced the number of significant descriptors for different regression models, namely PLS, ANN, RF and SVM.

Filter methods are independent of any predictive algorithm used in the later stage, hence they are typically used as preprocessing tools, whereas wrappers select features that specifically improve the accuracy of the classifier or regressor. Embedded methods combine characteristics from filters and wrappers and are implemented in algorithms which have their own built-in feature selection methods. [97]

Principal Component Analysis (PCA) is a feature extraction algorithm widely use for dimensionality reduction purposes. The algorithm uses orthogonal transformations, as depicted in Figure 3.14, to eliminate correlated features and projects the data onto a lower-dimensional hyperplane that preserves the maximum variance. Remaining independent variables are known as principal components. PCA is sensitive regarding the variance of the initial variables, hence, it requires data standardization to ensure every variable contributes equally to the analysis. [82, 84]

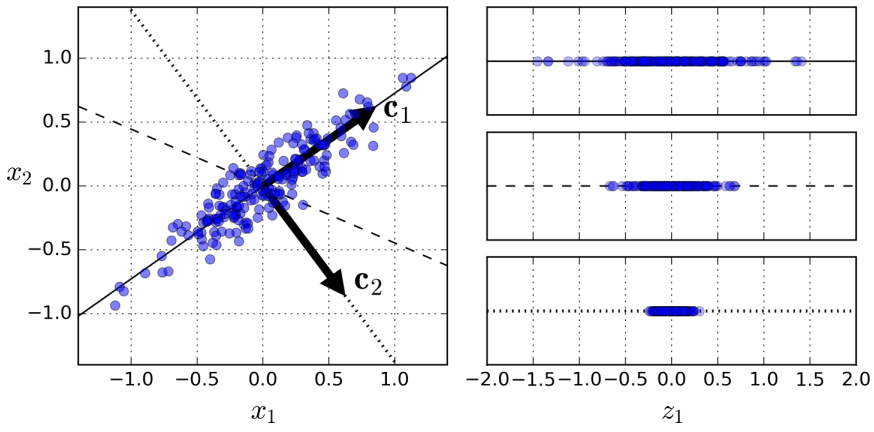

Figure 3.14 Selection of a lower dimensional space to project the data using PCA [82]

PCA analysis relies on linear transformations to reduce dimensionality, however, many real data sets are not linearly separable. Using kernel methods can improve the performance in those cases. [82]

4 Methodology

This section describes the methodology adopted in the present study. It is structured in five subsections that follow a logical order, from palette selection and data gathering to the actual modeling process.

4.1 Palette selection

Besides developing solid models for RON and MON prediction, this work aims as well for and accurate representation of gasoline and gasoline fractions. For such purpose, it is key to select the right palette of molecules, which must include not only hydrocarbons but also other frequent additives and property enhancers such as oxygenates. To be able to choose representative hydrocarbon species, the PIONA composition of typical streams blended into commercial gasoline (Figure 2.3) were carefully studied and one molecule was picked from each group. In addition to those, two typical octane boosters have been included, one alcohol and one ether compound. Data availability from existing publications has also played a decisive role in the palette definition.

The final compounds included in the palette together with necessary properties for the modeling process have been gathered in Table 4.1. RON and MON for n-heptane and iso-octane are well defined since they are used as reference fuels, however, significant variations are reported by different authors for the rest of the molecules. Hence, values proposed by Ghosh et al. in [40] were taken as a reference in this work except for 1-hexene [98] and ETBE [99]. Reference [100] served as a source for molecular weights and densities.

Table 4.1 Molecular palette selected for the construction of the predictive models [40, 98–100]

Molecule	**Group**	**Chemical formula**	**Molecular weight [g/mol]**	**Density [g/mL]**	**RON**	**MON**
n-heptane	n-paraffin	C_7H_{16}	100.21	0.6795	0	0
iso-octane	iso-paraffin	C_8H_{18}	114.23	0.6986	100	100
1-hexene	olefin	C_6H_{12}	84.16	0.6731	73.6	64.5
cyclopentane	naphthene	C_5H_{10}	70.13	0.7457	100	84.9
toluene	aromatic	C_7H_8	92.14	0.8623	118	103.5
ethanol	alcohol	C_2H_5OH	46.07	0.7852	108	92.9
ETBE	ether	$C_6H_{14}O$	102.17	0.7364	117	101

N-paraffins

The selected straight chain paraffin for the models was n-heptane. This molecule is frequently found in literature studies as part of simplified models [17]. Moreover, n-heptane is one of the reference fuels for octane number measurements with RON and MON values equal to zero. Although its presence is undesirable in fuels for SI engines due to its high knocking tendency, commercial gasolines typically contain 1 to 2% of n-heptane [16, 101].

Iso-paraffins

For many purposes, gasoline is simplified as a single molecule, iso-octane. Iso-octane is the second reference fuel used together with n-heptane for RON and MON measurements. Blends with high concentration of iso-octane are a close representation of high octane and low sensitivity streams such as alkylates [22].

Olefins

The percentage of olefins in gasoline is relatively low [17]. However, olefins are involved in important stages of gasoline refining. Olefins are produced in the FCC unit and combined with isobutene in alkylation units [22]. Moreover, due to the negative impact on gasoline properties [17] it is important to understand the interaction between olefins and other molecules to safely limit its presence in commercial products. The selection of 1-hexene in this work was motivated by data availability. Initially, 1-pentene was also considered.

Naphthenes

Like olefins, naphthenes are not found in large concentrations in gasoline [17], hence the limited impact of these components on its properties. Even though the presence of naphthenes in the final product is small, they play a significant role in the refining process. They are fed into the naphtha reformer together with a catalyst to produce reformate gasoline and they are also desired FCC feedstock [22]. In this paper, cyclopentane has been considered as a representative molecule.

Aromatics

Many ternary blends for gasoline representation consist of a mixture of the two primary reference fuels (iso-octane and n-heptane) and toluene. A blend of these three molecules in the right proportions can match the RON and MON of any given fuel [102]. Aromatics are characterized by their high octane rating and their presence in gasoline blends improves auto-ignition resistance [17]. While blends with high percentages of aromatics fairly represent streams such as gasoline reformate, mixture with absence of them can mimic the behavior of alkylates [22].

Alcohols

Ethanol is normally blended in commercial gasolines to enhance product properties and is preferred over methanol due to its lower toxicity. [17]

Ethers

An alternative to the use of alcohols in gasoline is the addition of ethers. ETBE has been selected over MTBE due to its increasing use [23].

4.2 Python and Scikit-learn

All the models included in this thesis have been created using Python and some of its most popular libraries. Python is a high-level programming language characterized by code readability and simplicity. Moreover, Python is the preferred language for artificial intelligence and ML applications [103]. Scikit-learn is the most used machine learning library available for Python. It provides a wide selection of supervised and unsupervised algorithms with the characteristic simplicity of the Python language [104].

4.3 Data collection, processing and splitting

Collecting the necessary data to train and test the algorithms was the most time consuming task of this master's thesis. Data collection is a tedious and slow process. However, it is crucial to guarantee the quality of the models since a predictive model is just as good as the data fed to it. Although the data included in the final database was obtained from 16 publications, it was necessary to examine substantially more sources. Studies often emulate gasoline by blending a limited number of compounds. However, each author selects the molecules that considers more beneficial and relevant for the purpose of their research. Despite some molecules being more common than others, it is challenging to find enough reliable sources using the same exact palette of species. Moreover, only experimental data has been included in the database and any computational results obtained from existing models or correlations have been ignored. The complete database can be found in Appendix 1.

The collected database was exported as a CSV file to be read using Python's pandas library. The result is a matrix of size 268×11. Columns 1 to 7 contain data regarding composition of each sample, column 8 refers to the number of components of the blends and columns 9 to 11 correspond to the values of the target properties, RON, MON and S. The number of rows exceeds in one unit the number of collected samples as labels for each column are included. Models have been trained and tested both for molar and volumetric data. For that, the database in Appendix 1 was converted into molar fractions using densities and molar weights reported in Table 4.1.

The initial database included 267 items, but after the elimination of duplicates the number of samples got reduced to 243. Rows with the same composition but different values reported for the target properties are not considered as duplicates and are kept in the database. While every sample includes RON values, MON and S are not so commonly studied and only 173 rows include the value for these properties. After the elimination of duplicates, data is shuffled to ensure randomness.

Figure 4.1 gives an overview on how the collected data is distributed. It can be noticed that the most abundant species in the database are n-heptane, iso-octane, toluene and ethanol. On the other hand, blends containing 1-hexene, cyclopentane and ETBE are seldom reported in literature. From the upper plots for n-heptane, it is clearly visible how increasing the content of n-paraffins has a negative effect on RON and MON.

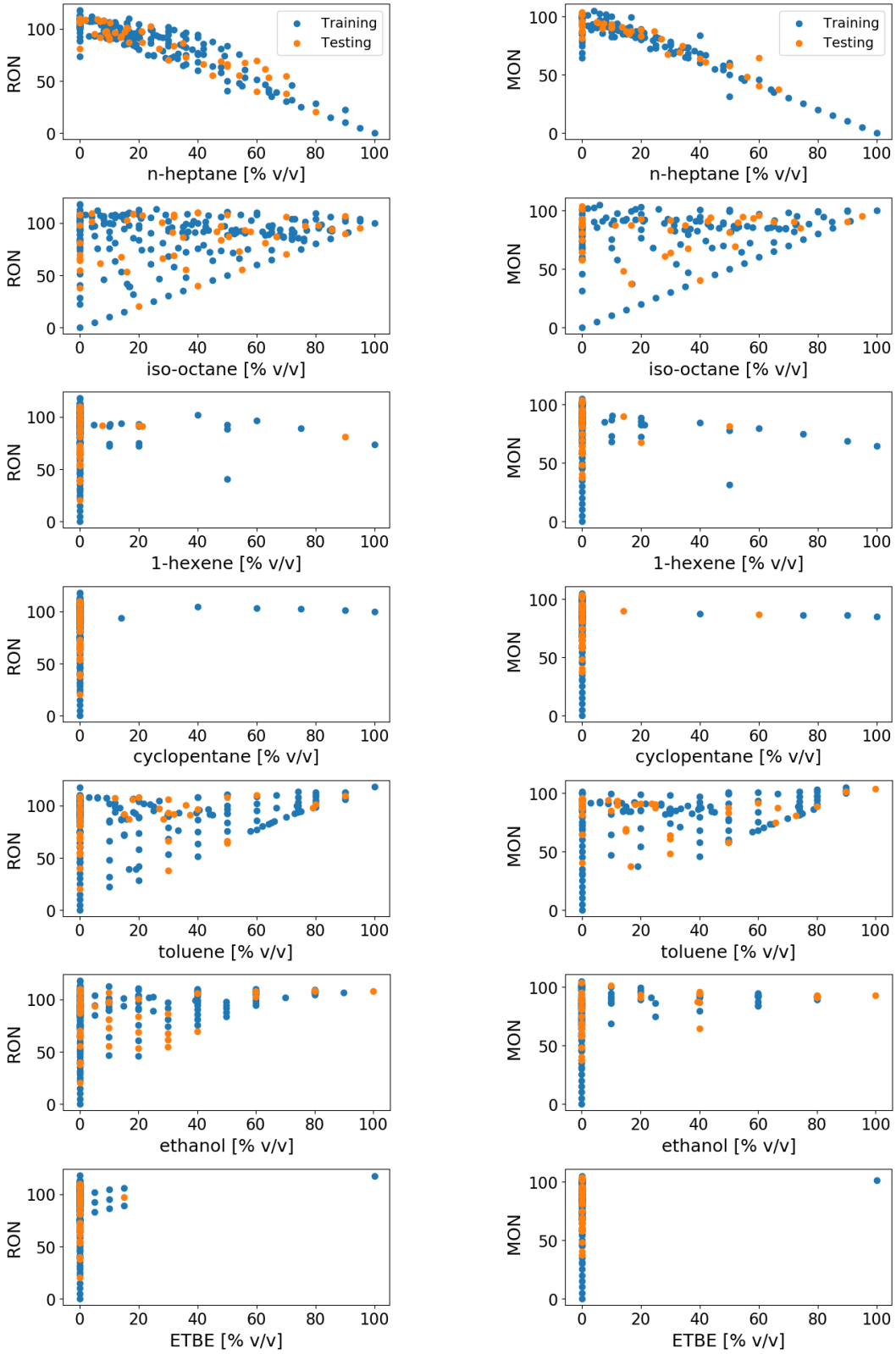

Figure 4.1 Sample distribution in the collected database for RON and MON

Figure 4.1 also reflects how the data has been divided into training set (blue markers) and testing set (orange markers). Building a predictive model typically involves three different stages, namely training, validation and testing, and each of these stages is associated with a different data set.

The training and testing data sets are always clearly differentiated. The training set is used during the training stage to fit the model, that is, to obtain the coefficients of a linear regression or the weights and bias for a neural network, for instance. On the other hand, the test data set is exclusively used once the model has been fine-tuned and trained to carry out an evaluation of its performance. This stage can also be seen as the external validation of the model.

The validation set affects the model in an indirect way. It offers an unbiased evaluation of the performance of the model to fine-tune its hyperparameters. The model is occasionally exposed to this data set but it does not technically learn from it. The way to select the validation data set will depend on the approach chosen for this stage. The traditional strategy reserves a portion of the original data exclusively for validation as shown in the upper part of Figure 4.2, which is seldom shown to the model during the training stage. However, when the available data is scarce, this reduces even more the actual number of training samples the algorithm can use to learn. Cross-validation overcomes that limitation by making use of dynamic subsets that change their status along the process. In this thesis simple k-fold cross-validation is used with k equal to 10, but other options such as leave-one-out are available.

Figure 4.2 Visualization of the data splitting for training, validation and testing according to different strategies

Cross-validation is interesting in this case due to relatively little data. Moreover, it allows for parameter fine-tuning and promotes model robustness. To perform k-fold cross-validation, the original data set is split only into training and testing sets in first place. Thereafter, the training set is further split into k subsets and the algorithm is trained k times. In each of the k folds, one subset is held out for validation while the remaining *k-1* subsets are combined as the training subset. This is shown in Figure 4.2 where k has been set to 10. In the first fold for example, the subset number 1 is selected as the validation subset while subsets 2 to 10 are used for training.

After training and validating the model k times, the performance over the different folds is compared. If there is considerable variation within the results, data might be too complicated for the chosen model or the algorithm might be unable to learn. However, if all folds return similar accuracy, it means the algorithm is consistent and it can be trained on all the data and externally validated using the test set [105].

All in all, there is no straightforward answer regarding how the data should be split between the three different sets, and the correct choice is very case dependent. In general, the training set is the largest one, accounting for 70 to 90% of the data. The remaining data is sacrificed for validation (if applicable) and testing purposes. Due to the use of cross-validation, all the models built in this thesis use 80% of the data for training and validation and 20% for testing.

4.4 Model selection

Section 3.2 presented the most popular supervised and unsupervised machine learning algorithms; however, not every algorithm suits every problem. Certainly, each algorithm possesses specific characteristics that can help to carry out a preliminary selection. In spite of this, finding the right approach requires of a trial and error process, where different models are fine-tuned, tested and later compared to each other.

Machine Learning algorithms can be divided into traditional methods and deep learning models. Deep learning algorithms, such as deep neural networks, improve their performance with increasing volumes of data. In contrast, when limited amounts of data are available, traditional and more simple algorithms show similar performance and they may even outperform deep learning models [106] as shown in Figure 4.3. For this reason, the scope of this work has been limited to traditional algorithms for regression analysis. Altogether, 8 ML algorithms are explored in this paper and their implementation is further explained in the next subsection. Those algorithms are Ordinary Least Squares (OLS), k-Nearest Neighbors (k-NN), radius-based Nearest Neighbors (r-NN), Linear Support Vector Regressor (LinSVR), Epsilon-Support Vector Regression (SVR), Nu-Support Vector Regression (NuSVR), Decision Trees (DT) and Random Forest (RF).

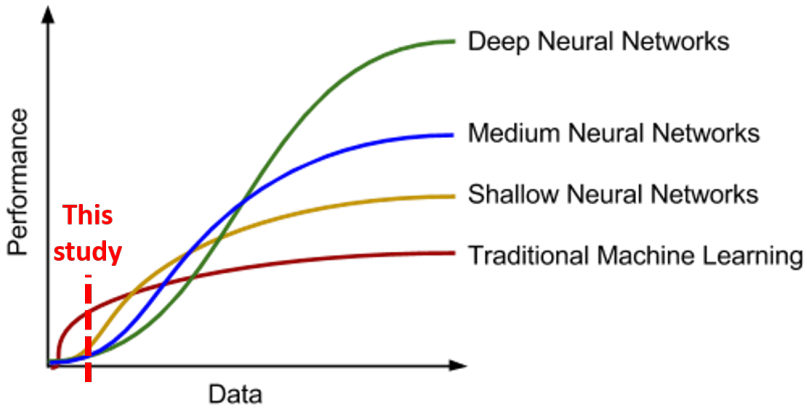

Figure 4.3 Performance of different ML techniques as a function of the amount of data available [106]

4.5 Modeling process

The pre-processed data as stated in section 4.3 was used to fit and compare the 8 chosen algorithms. Models include several parameters that can be adjusted to improve performance. Some parameters are common to most of them, such as normalization of the data or number of jobs used for the computation, while most of them are characteristic of each algorithm. Data standardization and normalization are common and recommended practices when building models. They refer to rescaling numerical attributes to have values between 0 and 1 and shifting numerical distributions to a mean of zero and a standard deviation of 1 to fit algorithm's requirements. However, do to the nature of the data used in this study, features can be already consider scaled while standardization was done separately during the training step. [85]

Relevant hyperparameters for each algorithm were tuned using Scikit-learn class `model_selection.GridSearchCV`. This class uses cross-validation techniques to find the optimal values for specified parameters for a given estimator. The cross validation technique can be defined by the user, and for this thesis simple 10-fold cross-validation was selected as already mentioned. The optimum is found using a scoring parameter that evaluates predictions on the test set, which was set to match model's score method. [85]

An important step when training any ML algorithm is the selection of a performance measure or cost function. It provides an estimation on the error the system is making during the prediction stage and allows for model comparison too. For regression tasks, it is common to choose Root Mean Squared Error (RMSE) as the cost function. In those cases where there are many outliers, the Mean Absolute Error (MAE) may provide superior results. During the training stage, models minimize the cost function using gradient descent optimization algorithm. Gradient descent enables the model to learn by "showing" the correct direction that minimizes the difference between predictions and targets. For simplicity reasons, default cost functions are used in the models unless otherwise specified. [82, 85]

The following subsections provide a more detailed explanation on how each algorithm was trained and internally validated, as well as which parameters were involved in the model tuning process. The main reference for these subsections is [85], where more detailed explanations can be founds for the different classes.

4.5.1 Ordinary Least Squares

Scikit-learn provides users with a wide set of methods to perform linear regression analysis, including OLS, Lasso, Ridge and Bayesian Regression, as well as variations and combinations of these methods [85]. Most of the algorithms utilize regularization techniques to reduce the number of coefficients based on multi-collinearities or include upgrades to deal with high-dimensional data. Nonetheless, those issues do not apply to the data in this work, hence OLS is sufficient to obtain satisfactory results.

Using Sckit-learn, OLS models can be developed using `LinearRegression` class included in `sklearn.linear` module. Parameter `fit_intercept` is set to `False`. [85] In this model, there is no need to calculate the intercept since the octane rating of a mixture is only dependent on its components, their concentrations and interactions among them. Remaining parameters are not specified and they take default values.

Simple k-fold cross-validation (`sklearn.model_selection.KFold` class) is used to validate the models, with number of splits set to 10 and remaining parameters taking default values [85]. Simple k-fold class with stated parametrization is used for cross-validation along the entire modeling process.

4.5.2 Nearest Neighbors

`sklearn.neighbors` module includes classes to implement Nearest Neighbors algorithms [85]. To perform regression analysis, it provides two alternative classes:

- `KNeighborsRegressor` averages the values of the k nearest neighbors of the query point [85].
- `RadiusNeighborsRegressor` learns from the instances within a circle of given radius r [85].

Model tuning is carried out using `sklearn.model_selection.GridSearchCV` class [85]. Table 4.2 includes the parameters and values tested for `KNeighborsRegressor` and `RadiusNeighborRegressor` respectively. The number of neighbors, and similarly the radius value, are very much dependent on the data set and type of problem. Therefore, a wide range of values was tested: from 1 to 10 neighbors and from 1.5 to 2.5 radius units with an increment step of 0.2 units. Uniform and proportional-to-distance weights are tested while euclidean and manhattan metrics are compared.

Table 4.2 Hyperparameter tuning for Nearest Neighbors algorithms

Parameters	Values	k-NN	r-NN
`n_neighbors`	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	✓	
`radius`	1.5, 1.7, 1.9, 2.1, 2.3, 2.5		✓
`weights`	uniform, distance	✓	✓
`metric`	euclidean, manhattan	✓	✓

4.5.3 Support Vector Machines

The `sklearn.svm module` provides three Suport Vector Regression methods. The `LinearSVR` class only considers linear kernels, `SVR` can utilize both linear and non-linear kernels and `NuSVR` permits control over the number of support vectors [85]. Table 4.3 shows the parameters evaluated for each algorithm and a description of them is also given next:

- `kernel` parameter defines which kernel function the algorithm is using. Two functions are considered besides the linear one: polynomic function of degree 3 and radial basis function. Each function has its own associated parameters. [85]
- `epsilon` defines the size of the epsilon-tube where no penalty is applied to the error of the predictions [85].
- `gamma` is a coefficient for non-linear kernels. Mathematically it is computed as the inverse of the standard deviation of the kernel function, and it defines the radius of influence of the samples selected as support vectors by the machine. Typical values range between 10^{3} and 10^{-3}. [85]
- `C` is a regularization parameter for the computation of the error term. Low values reduce the impact of missclasified samples or, in the case of a regression task, the impact of the prediction error. [85]
- `nu` coefficient sets a lower limit for the number of support vectors as a fraction of the available samples and replaces the parameter epsilon of the epsilon-SVR estimator [85].

Table 4.3 Hyperparameter tuning for SVM algorithms

Parameters	Values	SVR	NuSVR	LinSVR
`kernel`	rbf, poly	✓	✓	
`epsilon`	0, 0.01, 0.1, 0.5, 1, 2, 4	✓		✓
`gamma`	10, 1, 1e-1, 1e-2, 1e-3, 1e-4	✓	✓	
`C`	0.01, 0.1, 1, 10, 100, 1000, 10000	✓	✓	✓
`nu`	0.2, 0.4, 0.6, 0.8, 1		✓	

4.5.4 Decision Trees

`sklearn.tree` module and the class `DecisionTreeRegressor` are used to build Decision Trees from the input data [85]. Four parameters are tuned using grid search as shown in Table 4.4:

- `max_depth` sets the maximum depth that the tree is allowed to reach during the training stage. Small values limit the learning capacity of the tree, while too many levels lead to overfitting and large testing error. [85]
- `min_sample_split` sets the minimum number of samples in an internal node to be further splitted. [85]
- `splitter` can be set to `best` in order to split nodes based on the most relevant feature, or to `random` to make this decision in a random way. The latter may increase the complexity of the tree unnecessarily and reduce precision. [85]
- `criterion` defines the function to measure the quality of a split. By default, Mean Squared Error (`mse`) is used. `friedman_mse` uses Friedman's improve for potential splits, while `mae` uses Mean Absolute Error. [85]

Table 4.4 Hyperparameter tuning for Decission Tree algorithm

Parameters	Values
`max_depth`	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
`min_sample_split`	2, 3, 4, 5
`splitter`	best, random
`criterion`	mse, friedman_mse, mae

4.5.5 Random Forest

`RandomForestRegressor` class is an ensemble method fitting a certain number of Decision Trees to increase predictive accuracy and to reduce overfitting. Hyperparameter tuning of a random forest is very much linked to the tuning of the decision trees in it (`max_depth`, `min_sample_split`, `criterion`). In addition to that, the number of trees in the forest is controlled through `n_estimators`. [85] The tested values for each parameter are shown in Table 4.5.

Table 4.5 Hyperparameter tuning for Random Forest algorithm

Parameters	Values
`max_depth`	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
`min_sample_split`	2, 3, 4, 5
`criterion`	mse, mae
`n_estimators`	2, 4, 8, 16, 24, 32, 48, 64, 96, 128

4.6 Sensitivity analysis of the RON and MON models

To achieve a consistent database to train the models in this thesis, a few premises were established in the data collection stage. Samples taken into consideration must exclusively contain species included in the predefined palette (Table 4.1) and the values for RON, MON and S should always be experimental. However, other aspects such as the minimum or maximum number of components in the blends or thresholds for the octane numbers and octane sensitivity were not constrained.

From chapter 2.1 it is known that gasoline is a complex mixture of hundreds of different molecules. Moreover, octane rating of commercial gasoline must be high enough to guarantee normal combustion and engine durability. For these reasons, sensitivity analysis of the final RON and MON models was carried out by setting constrains in the data fed to the algorithms. Two parameters were adjusted to filter the data: octane number, either RON or MON, and complexity of the samples measured through the number of components in the blends.

Two approaches were taken regarding the octane number. Firstly, the database was filtered and only those samples over a certain lower boundary were retained. Selected values for that lower limit were 20, 40, 60, 80 and 100. In a similar way, the data was filtered a second time setting both upper and lower limits for the octane numbers. As a result, models were trained using samples contained in specific RON and MON intervals, in particular, [20, 40], [40, 60], [60, 80], [80, 100] and [100, 120].

A similar strategy was followed to analyze the effect of the number of blending components in the samples over the performance of the models. Smaller data sets were extracted from the original database where all the samples where binary, ternary or quaternary blends. In addition to that, subset with blends containing maximum 2, 3 or 4 molecules were also analyzed.

Following a similar methodology to the one used with the original models, the sensitivity subsets were split into training set with 80% of the samples, and test set with the remaining 20%. The cross-validation stage was neglected in this case to simplify the analysis, and the same hyperparameters used for the full database were kept.

5 Results and analysis

This chapter presents the results of the applied part of the thesis. Models are presented separately for three relevant auto-ignition properties related to the performance of gasoline fuels in SI engines. These properties are Research Octane Number (RON), Motor Octane Number (MON) and octane sensitivity (S), the later referring to the difference between RON and MON. Throughout the chapter the performance of eight different algorithms is analyzed and compared, both on the training and testing data sets.

5.1 RON models

RON, being one of the most important properties for commercial gasoline, must be carefully calculated according to the blending strategies of each refinery. The introduction of new molecules derived from alternative feedstock is a challenge for the existing models. Results presented in this section aim to bring new approaches to increase flexibility of the current tools.

5.1.1 Training and internal validation

The use of cross-validation techniques in this study serves three different purposes: model hyperparameter tuning, model evaluation and model comparison. Best performing hyperparameters for the 8 algorithms explored in this work were selected based on the results the 10-fold cross-validation. Moreover, the results of this stage are used for a rough estimation of the performance of the models on unseen data. Last, it provides with a solid basis to compare models before the testing phase.

The data used during this stage corresponds to 80% or 194 samples of the collected database, previously shuffled to ensure randomness and split in 10 consecutive subsets. For each fold, 9 of the splits were used to train the model and the remaining one was treated as a training subset for performance assessment. The results from the k-fold cross-validation for the RON models are shown in Table 5.1.

All models show a small standard deviation for the coefficient of determination R^2, which suggests data consistency and low risk of overfitting when training the final models. From Table 5.1 it can be noticed that using molar concentrations improved the performance on the training set in many cases. Nonetheless, the best performing model used SVR algorithm and volumetric concentrations.

Ordinary Least Squares

Volumetric and molar data yielded quite different results using OLS algorithm. When the model was trained with the volumetric data, the R^2 was 0.9163 and the average absolute error reached over 4 octane numbers. On the other hand, when the algorithm

Table 5.1 Cross-validation results for RON models

Model	Volume basis		Mole basis	
	R^2	MAE	R^2	MAE
OLS	0.9224 ± 0.0368	4.1592 ± 0.6421	0.9865 ± 0.0065	1.7685 ± 0.3365
k-NN	0.9647 ± 0.0250	2.8666 ± 0.8392	0.9555 ± 0.0392	3.0142 ± 0.9659
r-NN	0.4588 ± 0.0374	11.7828 ± 1.8965	0.5152 ± 0.0567	10.9722 ± 2.1312
LinSVR	0.9181 ± 0.0432	4.1893 ± 0.8035	0.9876 ± 0.0056	1.7409 ± 0.3268
SVR	0.9947 ± 0.0056	0.8973 ± 0.3622	0.9935 ± 0.0054	1.1445 ± 0.3556
NuSVR	0.9947 ± 0.0055	0.8913 ± 0.3604	0.9935 ± 0.0054	1.1402 ± 0.3525
DT	0.9395 ± 0.0253	3.6931 ± 0.6596	0.9685 ± 0.0177	2.7438 ± 0.7404
RF	0.9552 ± 0.0293	3.0542 ± 1.1354	0.9740 ± 0.0209	2.1456 ± 0.6764

was shown molar concentrations instead, the R^2 improved until 0.9809 and the absolute error dropped below 2 units. The resulting linear functions look as follows:

$$RON_{\text{vol}} = 11.5631 \cdot X_{\text{P}} + 101.0611 \cdot X_{\text{I}} + 73.0008 \cdot X_{\text{O}} + 95.1067 \cdot X_{\text{N}} + 117.1550 \cdot X_{\text{A}} + 127.6378 \cdot X_{\text{Ox}} + 117.5111 \cdot X_{\text{Ether}}$$

$$RON_{\text{mol}} = 1.4567 \cdot X_{\text{P}} + 102.5743 \cdot X_{\text{I}} + 73.2770 \cdot X_{\text{O}} + 97.5882 \cdot X_{\text{N}} + 114.4100 \cdot X_{\text{A}} + 114.2234 \cdot X_{\text{Ox}} + 117.5998 \cdot X_{\text{Ether}}$$

In general, the blending RON for n-heptane, iso-octane, ethanol and ETBE is higher than the actual RON of the molecules, while for 1-hexene, cyclopentane and toluene the situation is the opposite. However, there are big differences for some coefficients between both models, specially in the case of ethanol. To be able to capture the known boosting effect of ethanol in gasoline blends, the volumetric model allocates to this molecule a coefficient with a value as high as 127.45, far from the RON of the neat component ($\text{RON}_{\text{Ethanol}}$=108). Conversely, the coefficient in the case of the molar model is only 114.13, which suggests that ethanol blending behavior can be better explained on a mole basis. In fact, some existing publications [74, 107, 108] suggest that molar concentrations may be more appropriate to describe RON and MON dependence on alcohol content, and even on other oxygenated compounds such as ETBE and MTBE.

Considering that fuels are vaporized in the carburator of the CFR engine during RON and MON tests and that they are present in a gaseous phase in the cylinder, this behavior seems reasonable. Auto-ignition reaction rates are determined by the partial pressures of the gases involved, which according to the ideal gas law correlate linearly with molar concentrations. Furthermore, the liquid molar-volume ratio, determined using molecular weights and densities of the base fuel and the alcohol, explains why non-linearities in volumetric blending are more pronounced when molecules greatly differ in size. [107]

Using a molar approach for the blending problem exposes other phenomenon that do not arise using volumetric concentrations. When ethanol is blended in high-octane gasoline, the RON of the mixture increases linearly up to a certain value — several studies agree that this happens around 102-103 RON and variable concentrations — where it plateaus out. [107]

With second generation feedstock gaining attention, it is expected that other alcohols different from ethanol will be blended into petroleum products. Longer-chain molecules such as butanol and iso-butanol could be present in future gasoline, hence the importance of gaining a better understanding regarding the explained phenomena. [109]

Nearest Neighbors

Parameter estimation was carried out using grid search with k-fold cross-validation. Top performing parameters for both k-Nearest Neighbors and radius-based Nearest Neighbors are listed in Table 5.2.

Table 5.2 Top performing parameters for Nearest Neighbors algorithms in RON models

	Volume basis		**Mole basis**	
	k-NN	**r-NN**	**k-NN**	**r-NN**
`n_neighbors`	7	N/A	5	N/A
`radius`	N/A	1.9	N/A	1.9
`weights`	distance	distance	distance	distance
`metric`	manhattan	manhattan	euclidean	manhattan

Best results for k-NN algorithm were obtained with the number of neighbors equal to 7 and 5 for volumetric and molar data, respectively, and weights inversely proportional to the distance to the query point. The cross-validation results suggest a model improvement with respect to OLS algorithm, with R^2 over 95%.

For r-NN algorithm, distance-proportional weights were also chosen, with radius value equal to 1.9. In fact, smaller radii than 1.9 returned an error due to data sparsity. For some training points the algorithm could not find any neighbor, thus not being able to compute any valid output. In this case, Manhattan metric was prioritize over Euclidean distance. Nonetheless, the results are far from acceptable, with mean absolute errors larger than 10 octane numbers and poor coefficients of determination.

Distance-based weights prioritize the contribution of closer neighbors to the predictions. This behavior could be expected since closer neighbors have a similar composition, thus similar octane numbers. Manhattan metric accounts for the distance in every dimension, rather than considering just a straight line and typically yields better results for high-dimensional spaces. Unlike Euclidean distance, Manhattan approach provides with a better discrimination of the data when the ratio of the distances of the nearest and farthest neighbors to a given target gets close to 1 [110].

Support Vector Machines

Results from model fine-tuning for the different support vector regression algorithms are summarized in Table 5.3.

LinSVR did not improve the results of the OLS model in the volume basis and showed the lowest performance among all SVM algorithms. Moreover, the hyperparameter tuning process returned `epsilon` = 2 and `C` = 10000, which might cause some overfitting in the model. However, for the molar data satisfactory results were achieved despite

Table 5.3 Top performing parameters for SVM algorithms in RON models

	Volume basis			Mole basis		
	LinSVR	**SVR**	**NuSVR**	**LinSVR**	**SVR**	**NuSVR**
`kernel`	N/A	rbf	rbf	N/A	rbf	rbf
`gamma`	N/A	0.1	0.1	N/A	0.1	0.1
`C`	10000	10000	10000	10000	1000	10000
`epsilon`	2	0.5	N/A	2	1	N/A
`nu`	N/A	N/A	0.5	N/A	N/A	0.3

this being a linear algorithm. As for the OLS algorithm, this performance suggests that using molar compositions instead of volumetric measurements is a more effective way of explaining the blending behavior between different gasoline components [74, 107, 108].

For SVR and NuSVR, the preferred kernel function was the radial function in each and every case combined with large values for the penalty term `C`. In many cases, RBF tends to outperform other kernel functions due to lower tendency to overfit the model. However, large values of `C` might push in the opposite direction, since wrong predictions are highly penalized. Nonetheless, reflecting back on the numerical results in Table 5.1, it seems that a trade-off is achieved by combining both parameters, with R^2 exceeding 0.99 in all cases and volume-based SVR becoming the best performing model at this stage with MAE below 0.9 octane numbers.

Decision Trees

Table 5.4 shows the training parameters for the Decision Tree estimator. The DT for the volumetric data is trained to reach a depth of 17 levels using MSE as the quality indicator. The molar model is kept slightly simpler with 15 levels and quality of the splits is measured using MSE as well. The number of samples to further split an internal node is kept low, with 2 or 3 samples, most likely due to the relatively small training data set available. Last, features for the splits are chosen on a random basis for both models.

Table 5.4 Top performing parameters for Decision Tree algorithm in RON models

Parameters	**Volume basis**	**Mole basis**
`max_depth`	17	15
`min_sample_split`	2	3
`splitter`	random	random
`criterion`	mse	mse

Figure 5.1 is an actual visualization of the four first levels of the resulting DT for the molar data. It can be seen that the three first splits are made based on the concentration of n-heptane of the samples. The next level already considers a wider number of features, namely n-heptane, iso-octane, 1-hexene and toluene content. Nodes are colored using a color scale that represents the output value of the regression, with darker tones corresponding to higher RON. This possibility of getting a graphical representation of

the model is one of the main advantages of DT, although complexity rapidly increases as the tree grows deeper.

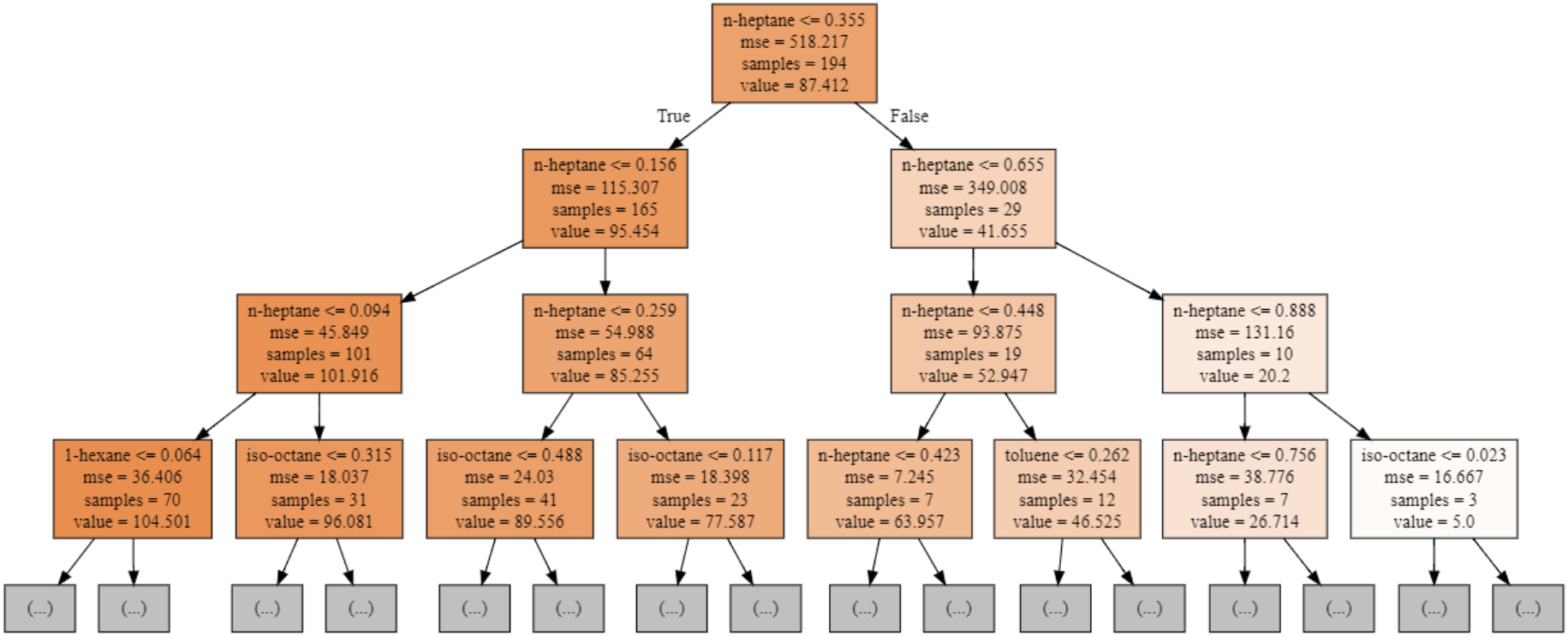

Figure 5.1 Visualization of the four first levels of the Decision Tree obtained for RON prediction on a mole basis

Random Forest

Trees grown by the Random Forest algorithm have a depth of 12 and 13 levels for volumetric and molar data, respectively. As it happened with the DT algorithm, the minimum number of samples to split an internal node is kept low and equal to 2, while the quality of the splits is measured using either MAE or MSE. As for the number of estimators, 24 was found to be enough for volumetric data, while 128 gave better results in the mole based model. These results are summarized in Table 5.5.

Table 5.5 Top performing parameters for Random Forest algorithm in RON models

Parameters	**Volume basis**	**Mole basis**
`max_depth`	12	13
`min_sample_split`	2	2
`criterion`	mse	mae
`n_stimators`	24	128

Given the substantial difference in the number of estimators for the two models a deeper analysis was carried out. The performance of the models as a function of the number of estimators used for the RT algorithm can be seen in Figure 5.2. On a volume basis (blue data on the left), R^2 slowly decreases on the test subset after reaching 24 estimators, while in a mole basis (orange data on the right) it improves with growing number of trees, although the change is almost negligible for values highers than 48.

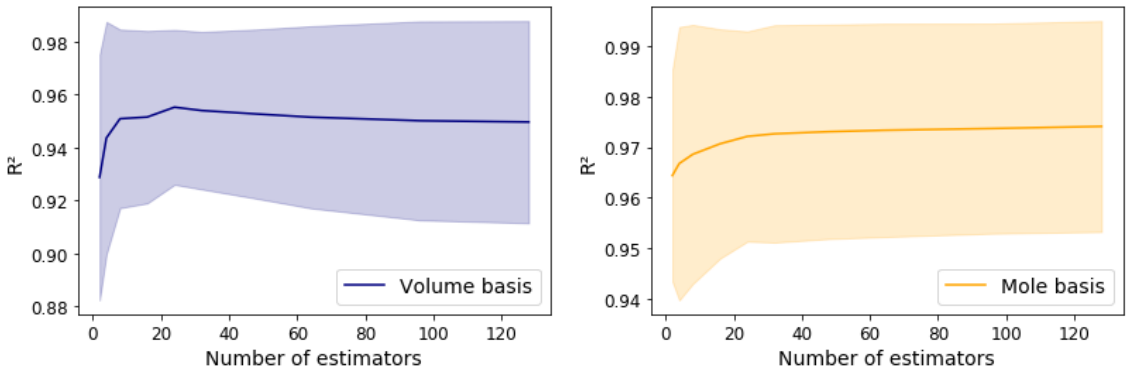

Figure 5.2 Impact of the number of estimators on the performance of the RON RF algorithm on a volume basis (left) versus on a mole basis (right) for 10-fold cross-validation

With small data sets like the one used in this thesis, using a large number of estimators does not cause any inconvenient. However, training models with larger volumes of data might result slow and require vast computational resources. In those cases, it should be considered sacrificing some accuracy and reducing the number of estimators of a RF model in order to speed up calculations.

5.1.2 Testing

After training the desired models with the best performing hyperparameters, the models were exposed to the test data set. In this case, the test set consisted on 20% of the collected database or 49 samples. The testing phase of the process is used to evaluate how well the algorithms have been trained and to compare the performance of the different models on unseen data. Moreover, this phase can be seen as an external validation of the models, although the data reserved for this step was acquired together with the testing data and do not have any characteristic features or distribution.

The performance for the 8 different models over the test set is presented in Table 5.6, and it can be seen that these results follow a similar trend to those from the cross-validation. Once more, SVR algorithm gives the best results, performing extremely good on a volume basis. The rest of the models, except for r-NN, also made accurate prediction on a mole basis and slightly worse with volumetric data.

Table 5.6 Performance of the trained RON models over the test set

Model	Volumetric data		Molar data	
	R^2	MAE	R^2	MAE
OLS	0.9255	3.9643	0.9884	1.7307
k-NN	0.9571	3.2107	0.9606	3.0234
r-NN	0.4891	12.1186	0.4766	12.3332
LinSVR	0.9180	4.1832	0.9880	1.7567
SVR	0.9962	0.9224	0.9903	1.4120
NuSVR	0.9964	0.9072	0.9903	1.4180
DT	0.8714	4.9806	0.9393	3.5622
RF	0.9701	2.6360	0.9852	2.0741

Some models achieved even better results on the test set than over the training set, as shown in Figure 5.3. Those models are OLS, r-NN, SVR, NuSVR and RF in volume basis, and OLS, k-NN, and RF in a mole basis. The only model showing a large gap between training and testing results is the volume based DT, which given the nature of the algorithm suggests overfitting.

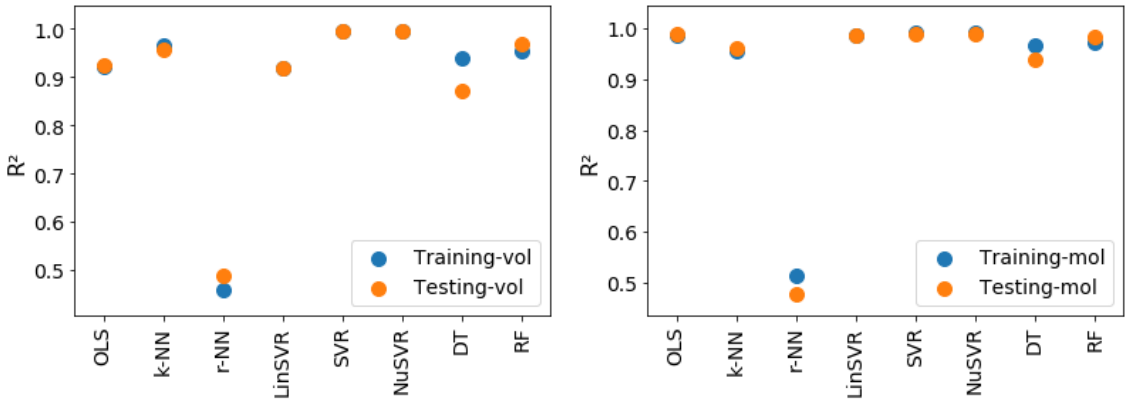

Figure 5.3 Average performance of the 8 algorithms over the training set versus the testing set for volumetric data (left) and molar data (right)

Figure 5.4 shows the predicted RON by the different models versus the experimental RON for all the points in the test set. It can be seen that most models performed better for higher RON values as more data was available for the the algorithms to learn in that region. Nevertheless, for too high RON values, the error starts to escalate again. This manifests the need for large amounts of properly distributed data to achieve well-trained unbiased models.

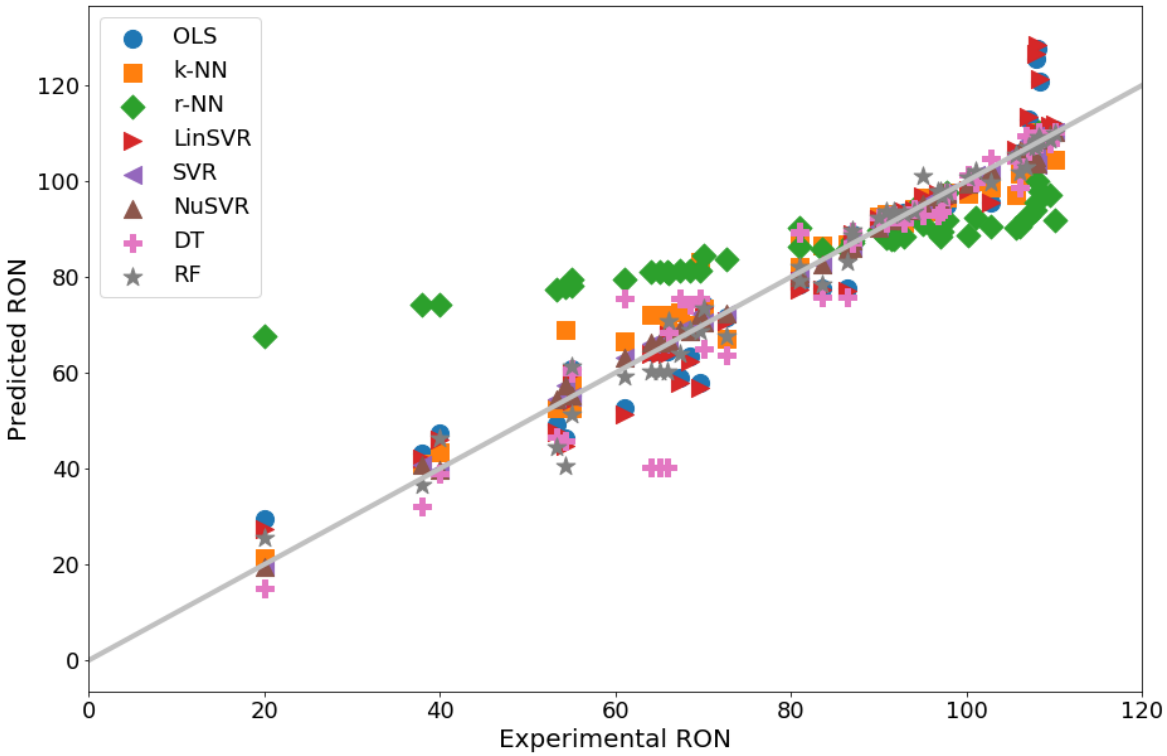

Figure 5.4 Predicted RON values for the samples in the test data set by the 8 trained algorithms versus the actual experimental RON for those points (volume basis)

From Figure 5.4 it might attract attention the inaccuracy of r-NN model, represented with the green markers. The reason for these results is explained by the mechanism of the algorithm itself and the distribution of the RON data. When r-NN model tries to make a new prediction, it computes an n-sphere[1] in the feature space of radius r around the query point. During the training of the model, hyperparameter r was assigned a high value to ensure that every query point had at least one neighbor within the domain of the n-sphere, even in those regions with low data density (i.e. low RON values). Despite the fact that this allows to predict RON for every sample, hinders high performance as neighbors are located too far and have too different composition. For dense areas, the issue is to some extent counterbalanced by a higher number of neighbors and the use of weights inversely proportional to the distance, which minimizes the impact of remote points.

[1]Term referring to the generalization of the ordinary sphere to spaces of arbitrary dimension.

5.1.3 Best performing model

After training and testing stages, SVR estimator showed the best performance among all the models, although the differences with several other algorithms were quite small as already pointed out in the previous sections. Accurate predictions for the SVR model confirm the suitability of the algorithm for relatively small but complex data sets as the one used in this study. Volumetric and molar data showed similar results, but the former scored slightly better on the training set and produced outstanding results on the test set. Nevertheless, these outcomes must be handle carefully as the same level of accuracy might not apply to future data sets.

Figure 5.5 shows the performance of both SVR models on the 49 samples reserved for the test set. The colors of the bars represent the composition of the samples on a volume basis and give information about the number of molecules in the blends, while the height represents the experimental RON of the sample. The dashed black line gives information about the volumetric model and its performance for each sample of the test set. The numerical values show those points where the error of the model was higher than two octane numbers.

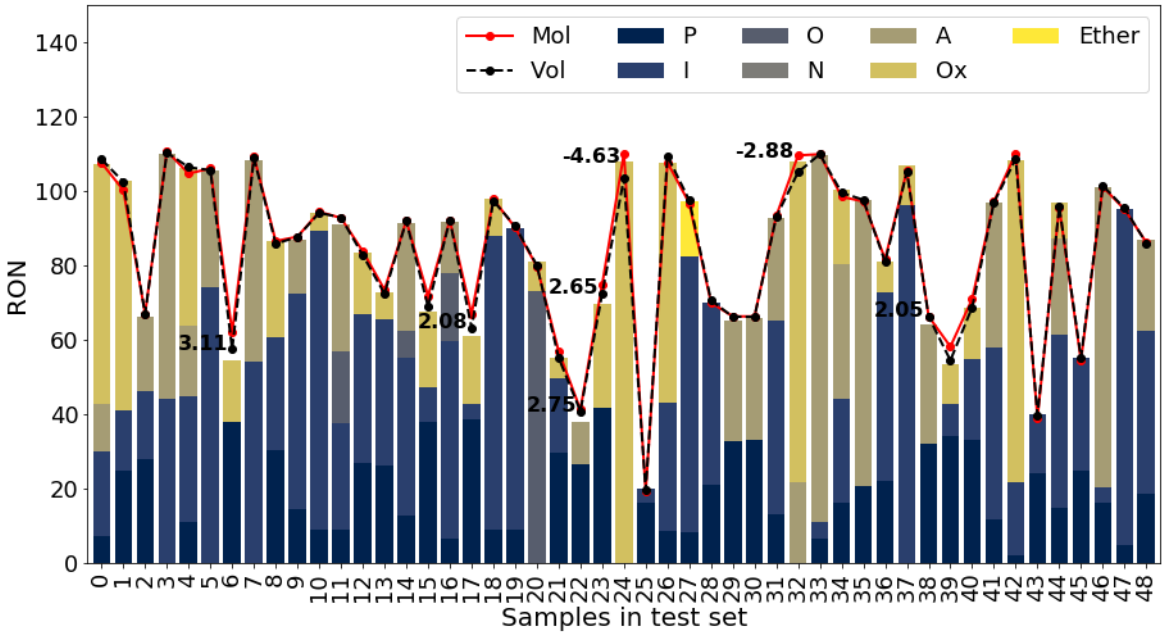

Figure 5.5 Performance of SVR algorithms over the RON test set, with values for absolute error exceeding two octane numbers included for the volumetric model

The large error of -4.63 octane numbers shown in Figure 5.5 for test sample number 24 corresponds to a composition of 100% ethanol. Given that most training points were blends of two or more molecules, it is reasonable that the model makes a wrong prediction for a pure compound. Moreover, it is worth to highlight that the model was never exposed to neat ethanol during the training stage as this sample belongs exclusively to the test set.

The rest of the errors reflected in Figure 5.5 reveal shortcomings on the database and set

guidelines for future research and model improvement. It can be seen that the predictions of the models are less accurate for samples with a low RON values or high concentrations of those species that were not so frequent in the database. Notwithstanding, RON for sample number 27, a blend of PRF90 with 15 %v/v ETBE, is predicted with high accuracy in spite of the limited number of ETBE-containing blends in the original data set.

Figure 5.6 and Figure 5.7 try to give an understanding of the learning level that the volumetric SVR model was capable to achieve. Figure 5.6 shows the blending behavior of binary blends of different hydrocarbon groups with ethanol according to the aforementioned model.

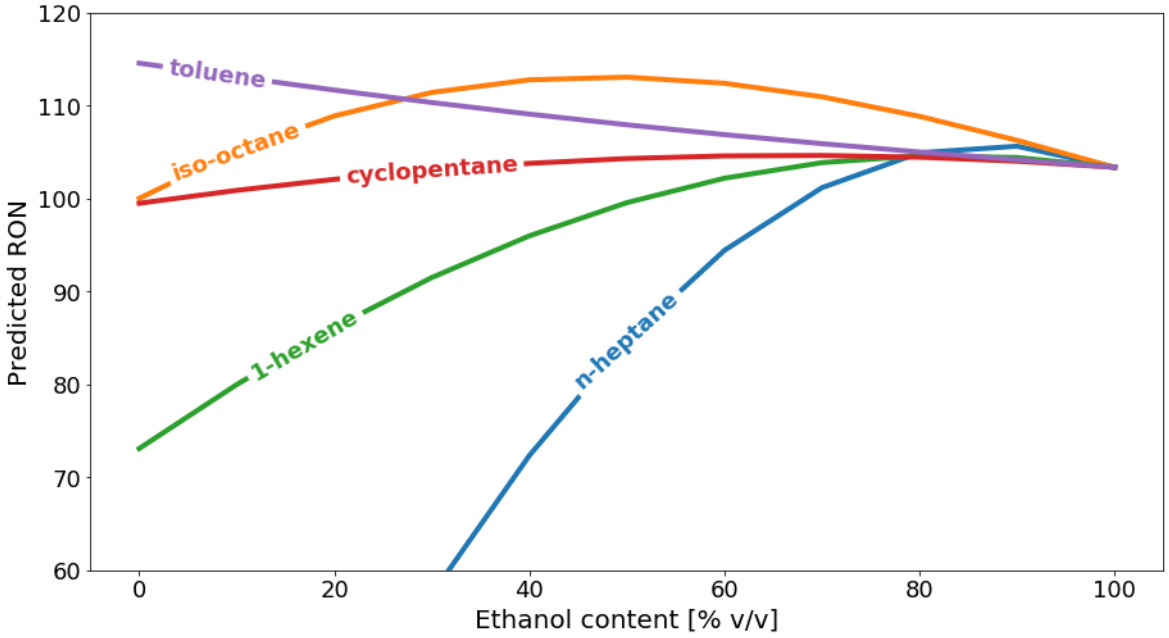

Figure 5.6 RON predictions for binary blends of ethanol and hydrocarbons using the volume-based SVR model

In Figure 5.6, n-heptane, iso-octane and 1-hexene blend synergistically with ethanol, although in the case of iso-octane the model exaggerates the effect showing hyperboosting betweeen 30% and 70% ethanol content. The synergism between the two paraffins in this study and ethanol is quite well known, in part due to these two molecules being used as PRFs. However, the same effect has been reported in literature for other species, such as n-pentane and iso-pentane [98].

The blending of cyclopentane and ethanol shows synergism to some extent. However, as explained for Figure 5.5 this model is unable to predict the RON of neat ethanol with accuracy. Hence the curvature downwards for the cyclopentane-ethanol blend in Figure 5.6, which otherwise should show neutral interaction [98]. For this same reason, the antagonistic blending effect for ethanol and toluene reflected by this model is not as strong as experimental data suggests [74]. This antagonistic blending has also been reported for other aromatic compounds such as 1,2,4-Trimethylbenzene both in mole and volume basis [98].

These blending behaviors for single hydrocarbons are very much in agreement with studies on real gasolines. A general finding is that gasolines with high paraffinic content show greater synergism when blended with ethanol than those with high concentration of aromatics [74]. This strong synergism with paraffinic gasolines is reflected by Figure 5.7, which shows the model results for ethanol blended in variable proportions with different PRFs, that is, mixtures of n-heptane and iso-octane. Once more, the model retains excessive synergism, reflected by the strong curvature of the lines, due to the wrong prediction for pure ethanol.

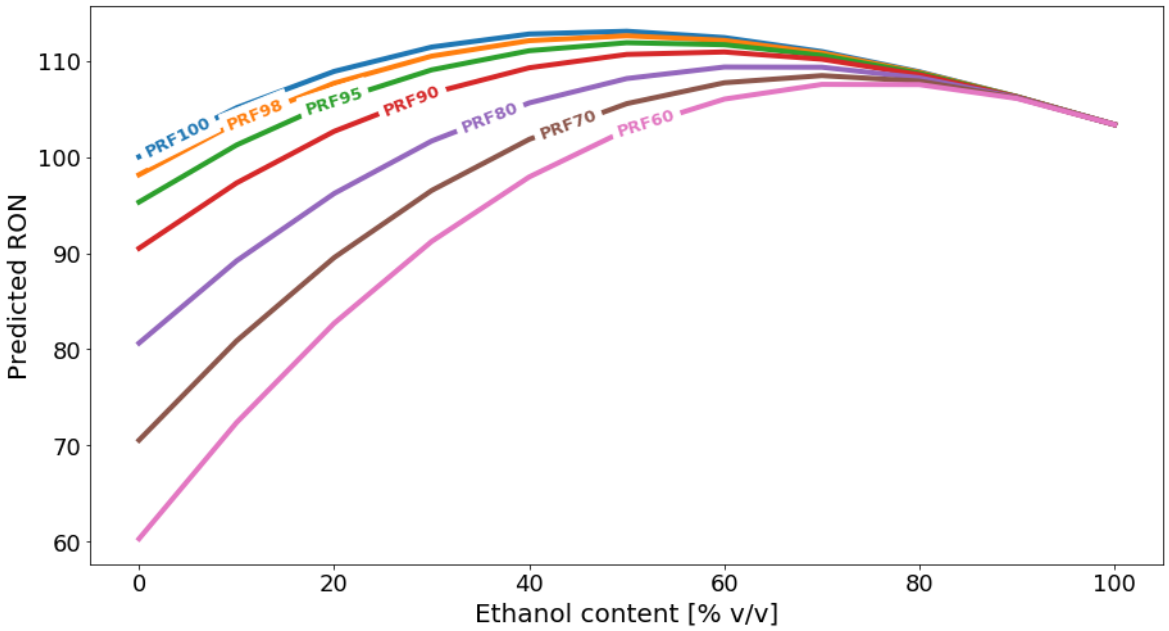

Figure 5.7 RON predictions for ternary blends of ethanol and PRFs using the volume-based SVR model

5.1.4 Sensitivity analysis of RON models

The sensitivity analysis of the fine-tuned models was carried out to understand the effect of the database structure and distribution on the predictive performance of the algorithms. In total, 16 different scenarios divided in four blocks were analyzed as stated in section 4.6, and three of the models which showed higher performance during the testing phase were selected for comparison. The intention of this analysis is also to see whether under different circumstances the performance of the SVR algorithm could be outperformed by other models, namely k-Nearest Neighbors and Random Forest. Despite giving accurate predictions, the NuSVR algorithm was not included in this analysis due to its resemblance with the SVR algorithm.

Figure 5.8 presents the results for the sensitivity analysis of the RON models. The first point in the data series corresponds to the reference case, which is the performance of the models on the original test set. Moreover, the size of the markers for each case reflects which portion of the database fulfilled the requirements.

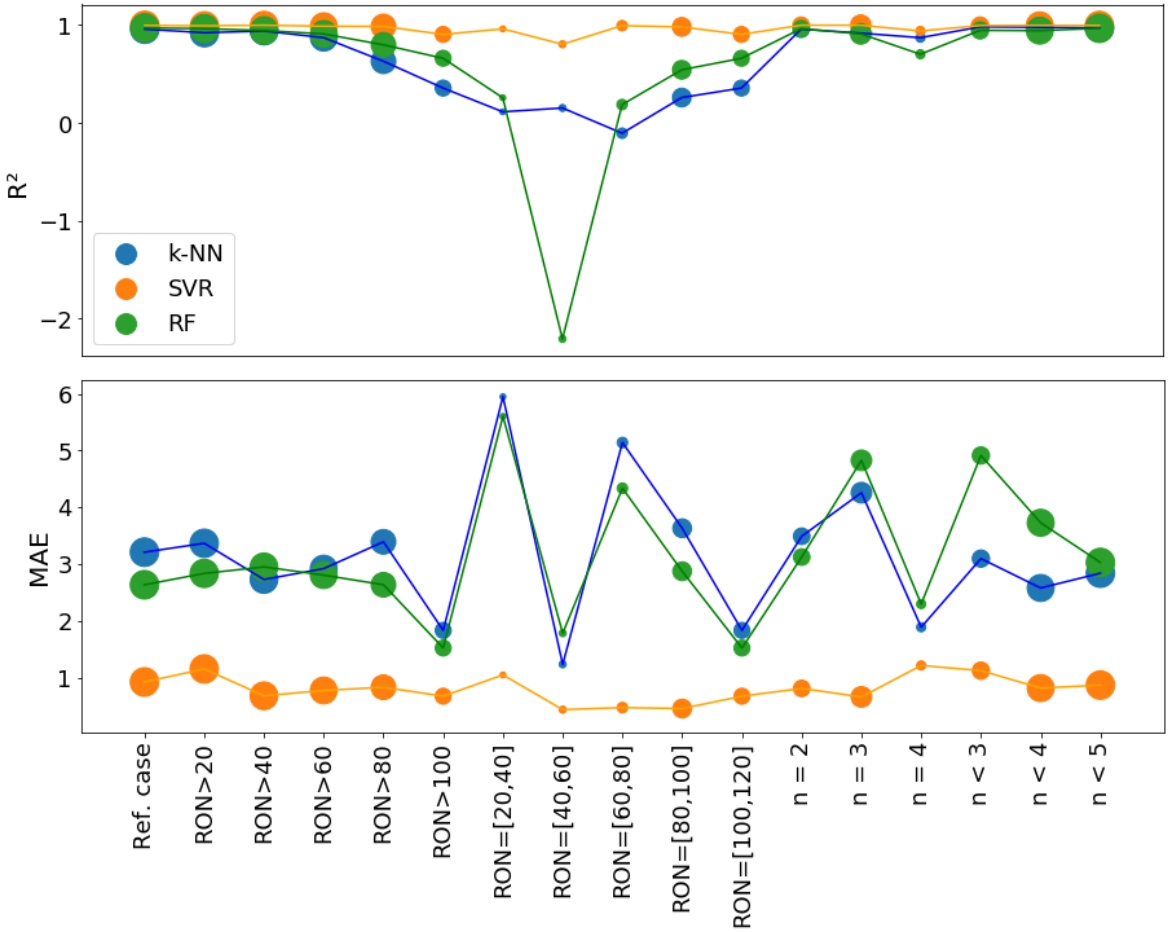

Figure 5.8 Sensitivity analysis of three RON models: k-NN, SVR and RF

When a lower boundary was set for the RON, the error made by the models experienced a general drop. Moreover, the lower availability of samples did not seem to have a large impact on the accuracy. However, when both an upper and lower limit were defined, models appear to suffer from this data deprivation, specially in the case of k-NN (blue data series) and RF (green data series).

In the seconds part of the sensitivity analysis, uncertainty was generated by constraining the number blended molecules in the samples taken from the database. Figure 4.6 shows large behavioral discrepancies among the three algorithms. For k-NN and RF, binary (n = 2) and ternary (n = 3) mixtures are more difficult to predict and the MAE of the models escalates up to 5 octane numbers, while quaternary models present lower error than the reference case. SVR model shows te opposite trend, with predictions in the quaternary model revealing lower accuracy.

Overall, the SVR algorithm proves to be more resilient than k-NN and RF to modifications in the distribution of the data set and its size. The coefficient of determination shows an stable balance for all sensitivity analysis cases and the mean absolute error never raised over 1.2 octane numbers.

5.2 MON models

Commercial gasolines in Europe are labeled according exclusively to their RON, however, quality standards include limits for MON too, as reported earlier in Table 2.1. MON captures the behavior of fuels in more harsh conditions than RON by using higher temperatures and rpms. While RON test settings simulate acceleration, engine running conditions during a MON test resemble fast speed driving for instance in a highway.

5.2.1 Training and validation

Similar to the RON models, MON models are evaluated during the training stage using 10-fold cross-validation both for hyperparameter tuning and model evaluation and comparison. Table 5.7 shows the average performance of the models over the training subset for the 10-fold cross-validation. Overall, 138 samples were used for training and cross-validation, which represents 80% of the available data, leaving the remaining 20% for testing purposes.

Comparing these result to those obtained for RON prediction, several algorithms here present higher standard deviation from the average performance for cross-validation, which suggests that the models are not able to generalize over unseen data. The reasons for that can be multiple: available data might be inconsistent, some of the folds might present different distribution or the nature of MON could be too complex for the chosen algorithms to learn.

Table 5.7 Cross-validation results for MON models

Model	Volume basis		Mole basis	
	R^2	MAE	R^2	MAE
OLS	0.8283 ± 0.3089	3.0823 ± 1.9481	0.8720 ± 0.3120	2.3180 ± 2.0747
k-NN	0.9639 ± 0.0290	2.1266 ± 0.5916	0.9472 ± 0.0337	2.6348 ± 0.7399
r-NN	0.4288 ± 0.1617	10.3843 ± 3.1954	0.4418 ± 0.1649	10.1687 ± 3.0720
LinSVR	0.8335 ± 0.3071	2.9576 ± 1.9502	0.8741 ± 0.3123	2.2392 ± 2.0760
SVR	0.9856 ± 0.0146	0.8964 ± 0.4820	0.9903 ± 0.0110	0.8514 ± 0.4288
NuSVR	0.9846 ± 0.0176	0.9183 ± 0.5172	0.9902 ± 0.0111	0.8676 ± 0.4404
DT	0.9526 ± 0.0306	2.8840 ± 1.1150	0.9674 ± 0.0200	2.3683 ± 0.9161
RF	0.9581 ± 0.0401	2.1266 ± 0.8601	0.9758 ± 0.0183	1.9435 ± 0.7945

Figure 5.9 shows how the test subset behaves for each fold of the cross-validation. It seems that 3 of the splits for the volumetric models and 2 for the molar data have a different distribution and do not respond well to the algorithms. Conversely, it can be noticed that the data series for the RON models are more compact, meaning that all cross-validation splits had similar characteristics and behavior.

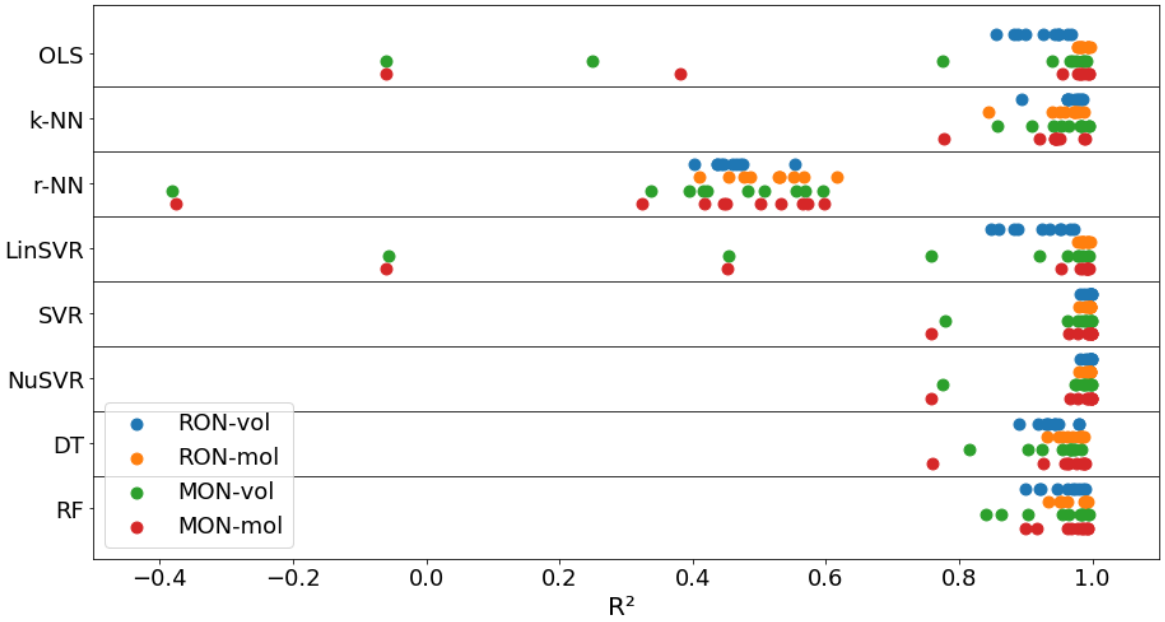

Figure 5.9 Comparison of the response of the training subset for the ten folds of the cross-validation process

Ordinary Least Squares

The functions resulting from the OLS algorithm are given below. Differences in the coefficients are significant between the two models, specially for n-heptane, toluene and ethanol. The molar model returns higher level of confidence with R^2 around 0.87 while the correspondent value for volumetric data is 0.82. This drop in accuracy with respect to RON models could be expected due to lower data availability, although the mean absolute error does not rise excessively, remaining around 3 octane numbers.

$$MON_{\text{vol}} = 4.7600 \cdot X_{\text{P}} + 98.1868 \cdot X_{\text{I}} + 63.6421 \cdot X_{\text{O}} + 83.4540 \cdot X_{\text{N}} + 106.8011 \cdot X_{\text{A}} + 97.5221 \cdot X_{\text{Ox}} + 101.0000 \cdot X_{\text{Ether}}$$

$$MON_{\text{mol}} = 1.12135 \cdot X_{\text{P}} + 101.1414 \cdot X_{\text{I}} + 61.2657 \cdot X_{\text{O}} + 83.7386 \cdot X_{\text{N}} + 102.0144 \cdot X_{\text{A}} + 93.8349 \cdot X_{\text{Ox}} + 101.0000 \cdot X_{\text{Ether}}$$

The regression coefficients for the mole-based model are closer to the MON of the neat molecules, suggesting that the blending behavior of the compounds might be better interpreted on a mole basis as discussed for the RON models and backed by previous studies [74, 107, 108].

Nearest Neighbors

Fine-tuning of the algorithm was carried out using grid search with 10-folds cross-validation also for MON models. Best performing parameters for both K-Nearest Neighbors and radius-based Nearest Neighbors are gathered in Table 5.8.

Best results for k-NN algorithm were obtained with the number of neighbors equal to 3 and 7 for volumetric and molar data respectively, Manhattan metric and weights

Table 5.8 Top performing parameters for Nearest Neighbors algorithms in MON models

	Volume basis		Mole basis	
	k-NN	**r-NN**	**k-NN**	**r-NN**
`n_neighbors`	3	N/A	7	N/A
`radius`	N/A	2.1	N/A	2.1
`weights`	distance	distance	distance	distance
`metric`	manhattan	manhattan	manhattan	manhattan

inversely proportional to the distance to the query point. For r-NN algorithm, distance-proportional weights and Manhattan metric were also chosen, with radius value equal to 2.1. Smaller radii than that returned an error due to data sparsity. For some training points the algorithm could not find any neighbor, hence not computing any valid output.

Support Vector Machines

Results from model fine-tuning for the different support vector regression algorithms are summarized in Table 5.9. The linear SVR algorithm shows just a slight improvement with respect to the results of the OLS model, while choosing a non linear kernel as RBF greatly boosts the accuracy of the predictions. The persistent high value for the penalty term C suggests that the models are minimizing the error on the training set, which may lead to overfiting and lower performance on unseen data.

Table 5.9 Top performing parameters for SVM algorithms in MON models

	Volume basis			Mole basis		
	LinSVR	**SVR**	**NuSVR**	**LinSVR**	**SVR**	**NuSVR**
`kernel`	N/A	rbf	rbf	N/A	rbf	rbf
`gamma`	N/A	1	1	N/A	1	1
`C`	10000	1000	10000	10000	10000	10000
`epsilon`	2	0	N/A	0.5	0.5	N/A
`nu`	N/A	N/A	0.3	N/A	N/A	0.2

Some similarities can be found with regard to RON models. First, the notorious improvement in the linear-based model when molar data is used instead of volumetric concentrations. Second, SVR and NuSVR behave in a similar way and yield comparable results as they just differ in one hyperparameter. Nonetheless, the NuSVR required longer training times which might bias model selection in future research on the topic if larger data sets are available. Last, these two algorithms showed the best overall performance during the training and cross-validation phase, reflecting the adaptability to relatively small and complex sets of data.

Decision Trees

Table 5.10 shows the training parameters for the Decision Tree estimators, which coincide for both molar and volumetric models. The grid search returns an optimal depth for the trees of 11 levels. Given that the number of features is the same as for RON models, the lower depths can be associated to the smaller data set available for the training in this case, although it might be also a consequence of the data distribution.

Feature selection for the splits is random, the minimum number of samples to proceed to split an internal node equals 2 and the performance measure is set to Friedman's MSE.

Table 5.10 Top performing parameters for Decision Tree algorithms in MON models

Parameters	Volume basis	Mole basis
`max_depth`	11	11
`min_sample_split`	2	2
`splitter`	random	random
`criterion`	friedman_mse	friedman_mse

The three upper levels of the obtained DT on a mole basis are shown in Figure 5.10. It can be noticed that this tree is simpler than RON DT presented earlier in Figure 5.1. Nonetheless, the split in the upper levels is done based on the n-heptane concentration for most nodes, with other species like 1-hexene or toluene, only introduced further down. The color of each box represents the value of MON: nodes on the left-hand side of the tree are colored in darker tones and lead to high MON values, while light-colored nodes on the right-hand side correspond with low MON samples.

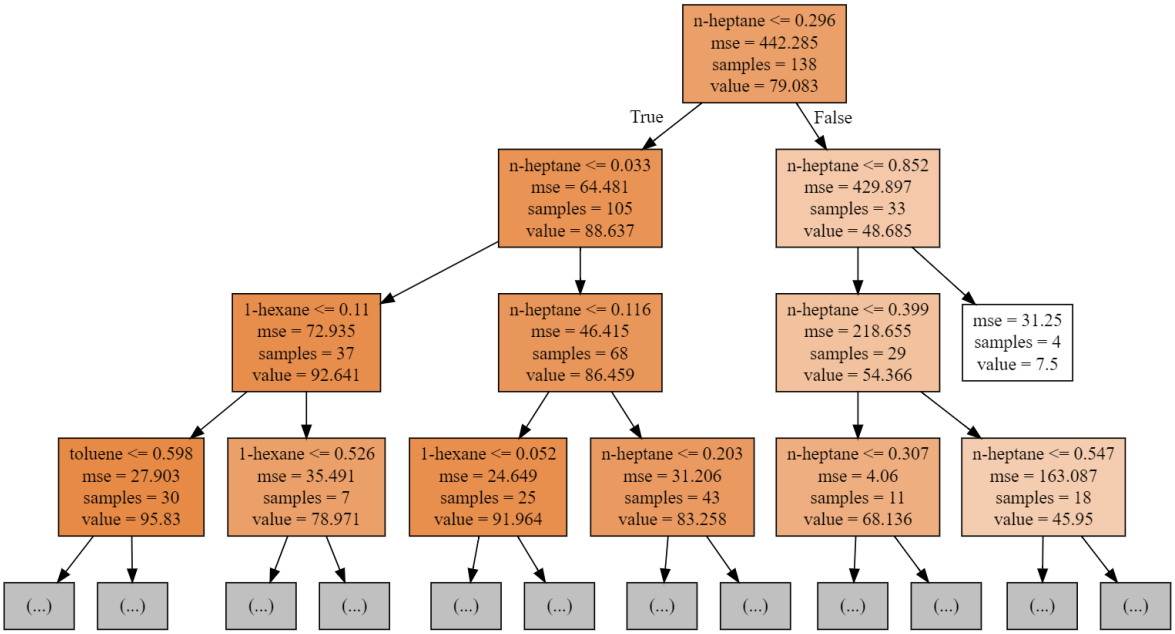

Figure 5.10 Visualization of the four first levels of the Decision Tree obtained for MON prediction on a mole basis

Random Forest

The Random Forest algorithm tries to increase the accuracy of the predictions made by the DT model by growing a larger number of trees and averaging their outputs. Table 5.11 shows the chosen hyperparameters for the RF estimator trained with MON data. Increasing the number of trees allows to slightly reduce their complexity by reducing the maximum depth to 10 levels for volumetric data and 9 for molar data, with he accuracy criterion set to MAE and MSE, respectively.

Table 5.11 Top performing parameters for Random Forest algorithms in MON models

Parameters	Volume basis	Mole basis
`max_depth`	10	9
`min_sample_split`	2	2
`criterion`	mae	mse
`n_stimators`	16	16

The number of estimators is set to 16 for both models. Intuitively one might think that the larger the number of estimators the better the predictions, at least in the training phase where the risk of overfitting is not so marked. However, as shown in Figure 5.11, performance can decrease when more trees are built by the algorithm as it might get contradictory information from different estimators. The negative impact of having excessive estimators is reflected by the solid line in the figures, which represents the average performance over the 10 folds of the cross-validation, but also by the colored area giving information about how differently the 10 folds reacted to the change in the number of trees.

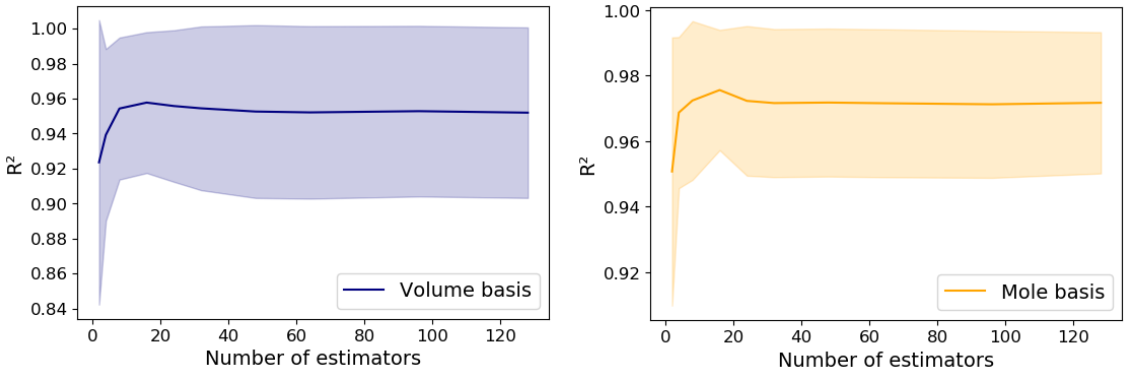

Figure 5.11 Impact of the number of estimators on the performance of the MON RF algorithm on a volume basis (left) versus on a mole basis (right) for 10-fold cross-validation

5.2.2 Testing

After the training and cross-validation steps, the fine-tuned models were exposed to the data reserved for testing purposes. The test data set is composed of 35 samples randomly chosen from the database which were not involved in the training process. The results of this stage are shown in Table 5.12 and help to identify possible overfitting and select the best performing models over unseen data. It can be seen that the performance of the different models is in accordance with the training stage. Again, SVR and NuSVR show consistency and excellent prediction capacity. Moreover, the two linear models OLS and NuSVR rank first on a mole basis.

Table 5.12 Performance of the trained MON models over the test set

Model	Volume basis		Mole basis	
	R^2	MAE	R^2	MAE
OLS	0.9105	2.8912	0.9838	1.6946
k-NN	0.9554	2.1970	0.8893	3.2332
r-NN	0.4644	9.9803	0.4507	10.0299
linSVR	0.9031	2.8942	0.9834	1.6847
SVR	0.9847	1.0534	0.9825	1.1670
NuSVR	0.9830	1.0667	0.9824	1.1804
DT	0.9086	3.2026	0.9607	2.2524
RF	0.8908	2.6531	0.9753	1.6206

The predictions provided by each molar model for the samples in the test set are shown in Figure 5.12. The closer the predictions are to the gray diagonal line the smaller the error made by the algorithm is.

The area below MON 40 does not contain any sample and the accuracy of the models in that region is unknown. Nonetheless, such low values are not so relevant for the scope of this study as most gasolines and gasoline streams tend to show higher antiknock quality, with MON 60 and superior. This is a direct consequence of the lack of samples in that area in the original database, and manifests that gathering more experimental data is needed if accuracy of prediction was required for low MON values.

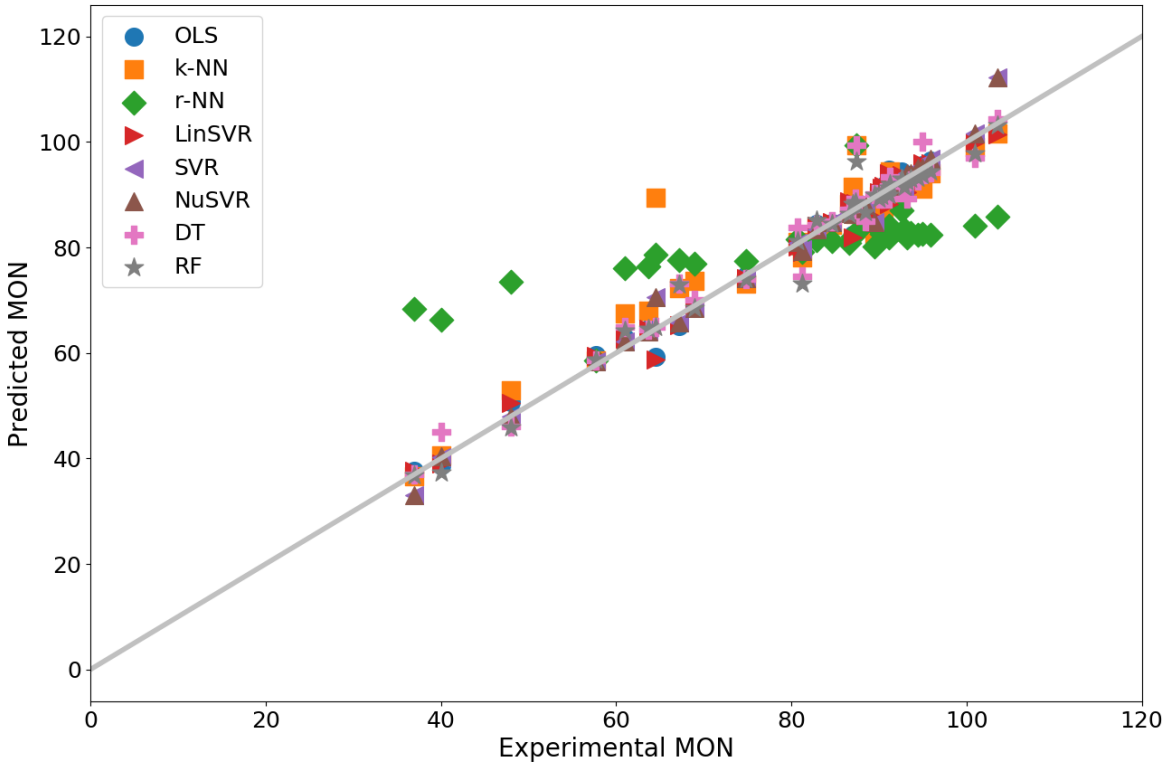

Figure 5.12 Predicted MON values for the samples in the test data set by the 8 trained algorithms versus the actual experimental MON for those points (mole basis)

5.2.3 Best performing model

The SVR algorithm trained with volumetric data was found to rank first on the test set with a value of R^2 equal to 0.9847 and MAE of roughly one octane number. However, the difference in performance with the molar OLS models is minimum, which shows an R^2 of 0.9838 despite a higher error of 1.6 units. The predictions made by these two models are shown in Figure 5.13, where the solid red line represents the performance of the OLS model and the dashed black line refers to the predictions of the SVR model. The color and height of the bars correspond to the composition and experimental MON of the samples respectively.

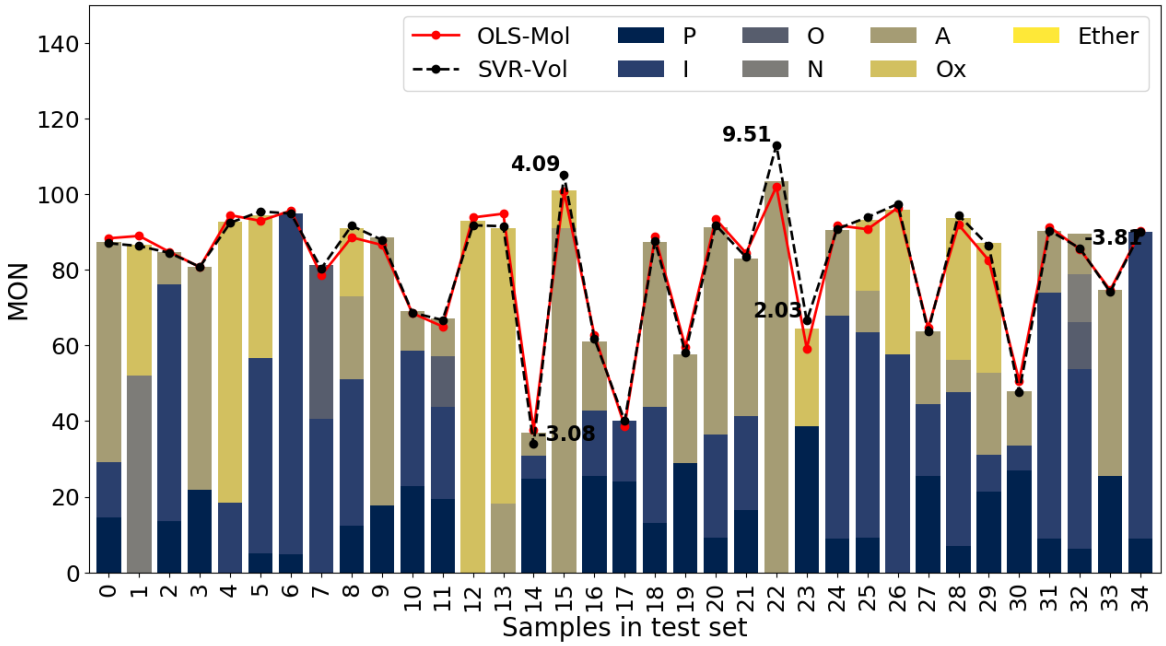

Figure 5.13 Performance of the two best models over the MON test set, with values for absolute error exceeding two octane numbers included for the volumetric SVR model

For the SVR algorithm, the model shows similar shortcomings to those of the RON model. The largest error in point 22 (9.51 MON) corresponds to a neat substance, in this case pure toluene. Also, sample number 15 with a composition of 85% toluene and 15% ethanol is predicted with a high deviation due to the low learning rate of the model in regions with high content of the aromatic hydrocarbon.

5.2.4 Sensitivity analysis of MON models

Sensitivity analysis of the best MON models was carried out following the same procedure as for RON models but two study cases sere skipped, MON $>$ 100 and MON $=$ [100, 120]. Since values for MON are generally smaller than RON vales, the number of samples fulfilling those conditions was too low and models provided inconsistent results. The scores for the remaining 14 cases are shown in Figure 5.14 for four different models:

volumetric k-NN, volumetric SVR, molar OLS and molar LinSVR. The sensitivity analysis for MON models is especially interesting since many models showed a similar performance when evaluated on the test set.

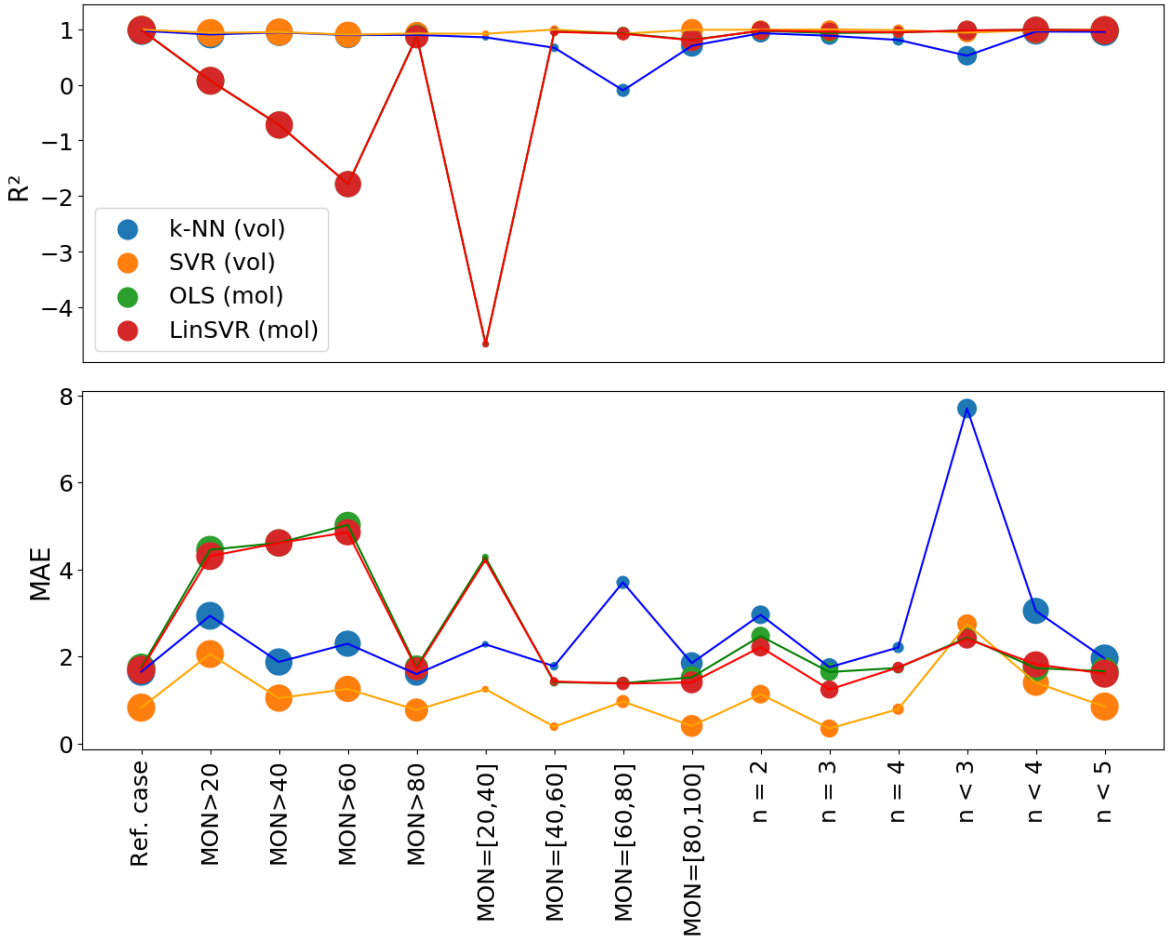

Figure 5.14 Sensitivity analysis of four MON models: k-NN and SVR on a volume basis and OLS and LinSVR on a mole basis

In the upper part of Figure 5.14 the value of R^2 is plotted for each sensitivity case study, while in the lower part the mean absolute error is presented. The size of the markers gives an estimation of how big the considered data sets were with respect to the original one. It can be noticed that setting a lower boundary for the MON values affected the accuracy of the models only for small values. However, when the models where only exposed to samples with MON $>$ 80 the error dropped for all algorithms. Similarly, when MON in the interval [80, 100] was selected as a filtering condition, all models improved their predictions. Having exclusively pure or binary samples (n $<$ 3) severely affected the quality of the results; moreover, it was the only case in which SVR model was outperformed by the rest of algorithms.

Overall, it can be said that the SVR algorithm restates its suitability for the prediction of octane numbers even in a modified or biased data environment.

5.3 S models

The octane sensitivity is defined as the difference between RON and MON for a given fuel. Therefore, the previously presented models for RON and MON could be used in combination to calculate the S of a new blend. This approach gives a new model, where the prediction error for S is equal to the summation of the RON error plus the MON error. For this reason, independent models were trained for S and later compared with the compounded model to investigate if the results could be improved.

5.3.1 Training, validation and testing

The modeling process for the octane sensitivity included the same steps as RON and MON modeling processes: training and cross-validation first and testing during a later stage.

The size of the available data set was 173 samples split in two groups, 80% for training and cross-validation and the remaining 20% dedicated for testing purposes. The results of the cross-validation, gathered in Table 5.13, show a general improvement in the predictions using molar compositions, although volumetric SVR gives the best training results. On a mole basis, k-NN, SVR, NuSVR and RF show similar performances with R^2 around 0.92.

Table 5.13 Cross-validation results for S models

Model	Volume basis		Mole basis	
	R^2	MAE	R^2	MAE
OLS	0.8191 ± 0.1449	1.4277 ± 0.4309	0.8554 ± 0.1478	1.1049 ± 0.3967
k-NN	0.8921 ± 0.1013	0.9412 ± 0.4016	0.9231 ± 0.0575	0.8118 ± 0.2613
r-NN	0.4334 ± 0.1017	2.9671 ± 0.8378	0.5118 ± 0.0900	2.7706 ± 0.8047
LinSVR	0.8189 ± 0.1493	1.3411 ± 0.4622	0.8578 ± 0.1492	1.0395 ± 0.4076
SVR	0.9307 ± 0.0451	0.8055 ± 0.2892	0.9282 ± 0.0437	0.8490 ± 0.2254
NuSVR	0.9160 ± 0.0374	0.9386 ± 0.2723	0.9190 ± 0.0442	0.8491 ± 0.2635
DT	0.8080 ± 0.1546	1.3173 ± 0.4356	0.8114 ± 0.2642	1.1574 ± 0.5358
RF	0.9038 ± 0.0663	1.0081 ± 0.3083	0.9191 ± 0.0501	0.9327 ± 0.3210

The selected hyperparameters for these five top performing algorithms are included in Table 5.14. The k-NN algorithm was trained with Manhattan metric, distance-based weights and a low number of neighbors, 4. These results can be again linked to the fact of having a high dimensional space and the similar behavior among samples with similar composition. The selected parameters for the two support vector regressors differ slightly, with a higher value for `C` in the case of SVR. Last, the RF algorithms in composed of 16 trees with a maximum depth of 14 levels. The minimum number of samples to split an internal node is set to 2 and MAE is used as the criterion to measure performance.

Table 5.14 Hyperparameter selection for S models showing high predictive accuracy in the trainig stage

k-NN	SVR	NuSVR	RF
`metric`: manhattan	`kernel`: rbf	`kernel`: rbf	`max_depth`: 14
`n_neighbors`: 4	`gamma`: 1	`gamma`: 1	`max_sample_split`: 2
`weights`: distance	`C`: 100	`C`: 10	`criterion`: mae
	`epsilon`: 0.5	`nu`: 0.6	`n_estimators`: 16

After fine-tuning and training the algorithms, they were tested over the unseen samples. The results of the testing stage are reported in Table 5.15. Despite the results in the cross-validation, the SVR model did not succed to predict the octane sensitivity for the test set. Instead, NuSVR model trained on molar data achieved the highest accuracy with mean average error below 0.8 octane numbers.

Table 5.15 Performance of the trained S models over the test set

Model	Volume basis		Mole basis	
	R^2	MAE	R^2	MAE
OLS	0.8303	1.6679	0.9004	1.1687
k-NN	0.8848	1.0841	0.8328	1.1191
r-NN	0.4151	3.2733	0.5255	2.9964
LinSVR	0.8378	1.5296	0.9084	1.0813
SVR	0.8860	0.8059	0.9084	0.8285
NuSVR	0.7927	1.0035	0.9415	0.7878
DT	0.7837	1.4788	0.7960	1.4618
RF	0.8572	1.2538	0.9099	0.9035

For an easier comprehension and visualization of these results, Figure 5.15 shows the values predicted by the 8 molar models versus the experimental value of S for the 35 samples included in the test set.

When compared to the analogous figures presented earlier for RON and MON models (Figure 5.4 and Figure 5.12) S predictions show higher dispersion. This is a symptom of general learning constrains coming from intrinsic characteristics of the data. Despite the data corresponding to the r-NN algorithm in green still standing out from the rest, results from other models such as k-NN (orange) or DT (pink) are quite scattered too.

Additionally, octane sensitivity does not show a predominant value or range of values as RON and MON do. As a consequence, there is no specific region where more training samples were concentrated making the models more accurate at least in certain spots.

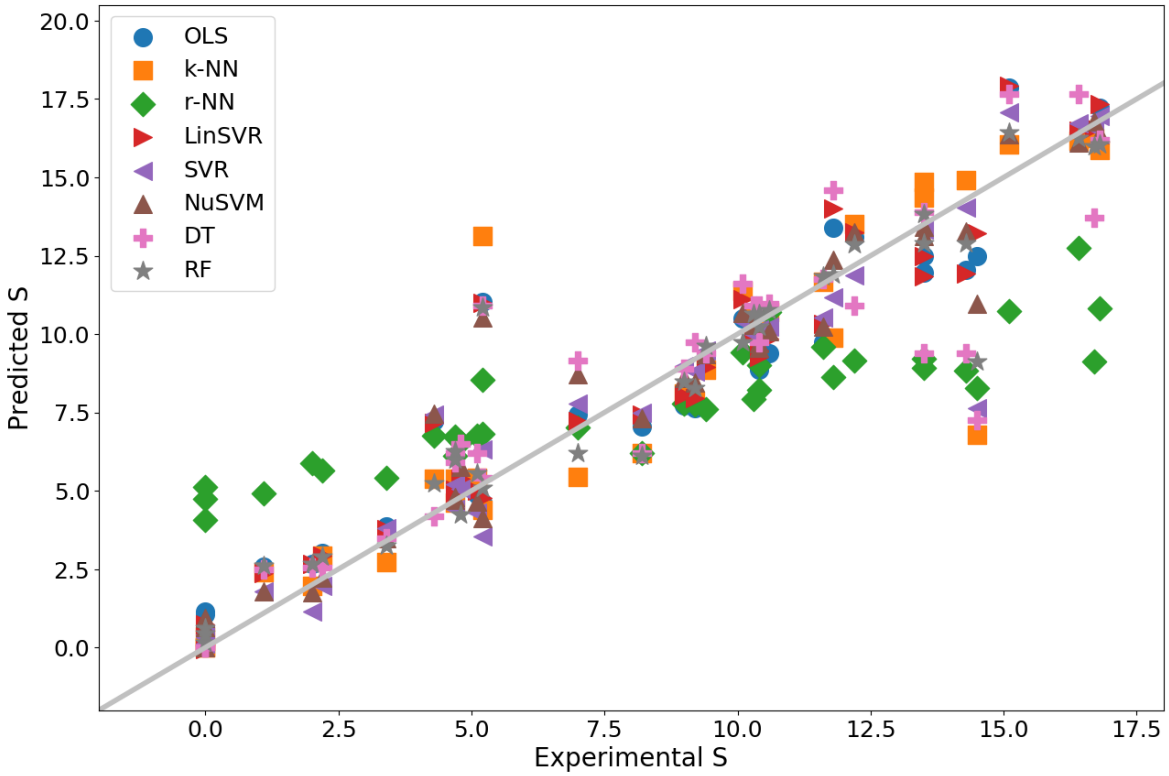

Figure 5.15 Predicted S values for the samples in the test data set by the eight trained algorithms versus the actual experimental S for those points (mole basis)

5.3.2 Simple model versus compounded model

After training and testing the simple S models, molar NuSVR was selected for its high performance to be compared with the compounded model. The compounded model was obtained as a combination of the best RON and MON models, where S values were obtained by subtracting the prediction of the volumetric SVR MON model from the volumetric SVR RON model.

The predictions of these two models for the test set are shown in Figure 5.16. The colors of the bars represent the composition of the samples on a volume basis. However, it must be noticed that only the compounded model works with volumetric data, while the simple model is based on molar compositions. The height of the bars is given as a reference and it represents the experimental S of the samples. The dashed black line gives information about the simple model and its performance for each sample in the test set. The solid red line follows the predictions from the combined model which, as expected, made larger errors than the simple model.

The numerical values corresponds to those points where the error of the simple model was higher than one octane number. Test points number 12 and number 22 belong to neat samples of ethanol and toluene respectively, hence the discrepancies between the model and the experimental values. Since the model never saw those species isolated during the training phase, difficulties to predict their octane sensitivity in the later

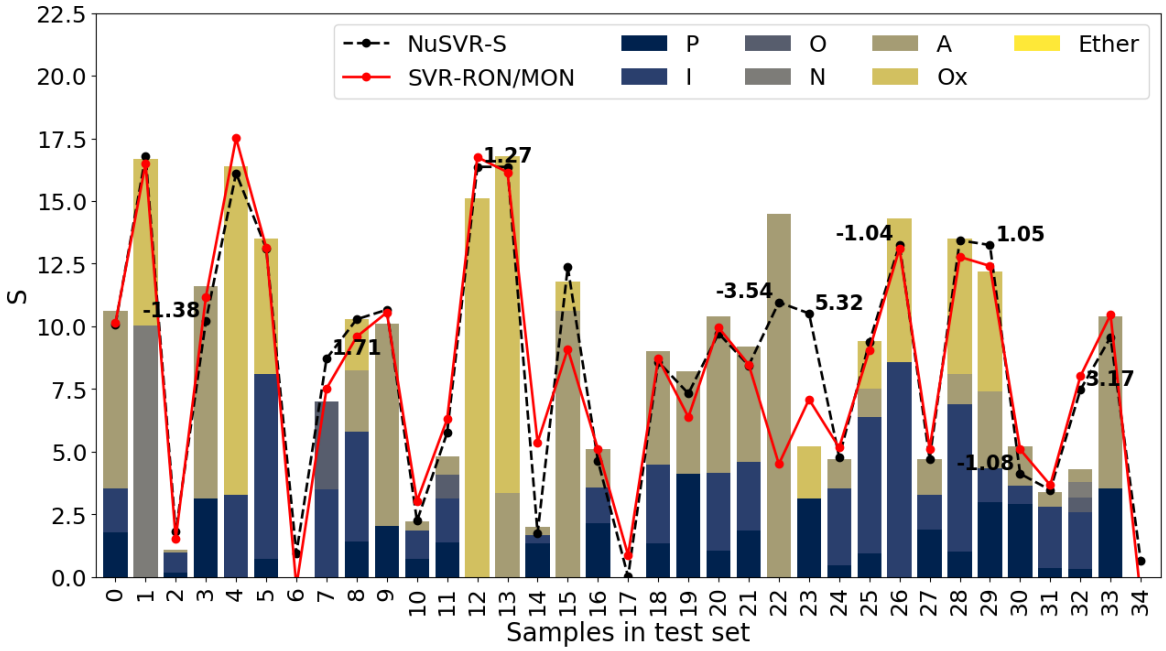

Figure 5.16 Predicting performance of the simple S model versus the compounded S model on the test set

stage emerge.

Although having separate models for the three properties gives the best results, in practice only two of them can be used, while the third property should be calculated in an indirect way from the other two to avoid inconsistencies. An approach that could potentially improve S predictions in the compound model would be to train RON and MON algorithms as multi-output models where S is computed in parallel.

6 Conclusions

This master's thesis addressed the question of fuel blend property prediction with special focus on fuels for spark ignition engines. Fossil-based gasoline is a complex mixture of several hundreds of molecules that interact with each other at different stages causing non-linear blending behaviors. With alternative raw materials for gasoline production gaining attention, the question of their impact on the properties of the end-product arises. In particular, non-linear interaction between gasoline molecules for auto-ignition properties has been suggested before in the literature, where two main behaviors are distinguished, namely synergistic and antagonistic blending. Simple linear models are unable to capture the nature of these phenomena and fail in making accurate predictions. In that context, this work explored more complex mathematical tools that can potentially give a better response to this problem.

To enable the study, gasoline composition was simplified and represented with a palette of seven chemical species, including five hydrocarbons, one alcohol and one ether. This approach responds to the expected variation in composition of future gasolines, with increasing contents of oxygenated compounds or at least a greater variety of molecule types. For the exact same reason, the scope of the analysis was not constrained to compositions resembling typical gasoline, but extended to any combination of the species in the palette with the aim of obtaining more flexible models. A database with 243 different blends of the selected molecules, expressed both in molar and volumetric concentrations, was collected from 16 different scientific publications along with corresponding experimental RON values. For 178 of those samples MON values were also found, and octane sensitivity was calculated in those cases. The resulting data set was divided into two subsets, one for training containing 80% of the samples and another one for testing with the remaining 20% of the data. Given the relatively small size of the data sets, simple 10-fold cross-validation was used for internal model evaluation and fine-tuning purposes.

Machine Learning was used to develop predictive models for the three auto-ignition properties RON, MON and S. Eight algorithms with variable levels of complexity were assessed. Linear algorithms like OLS and LinSVR lack mechanisms to reflect non-linear relationships within the data and incurred in large predictive errors when trained with volumetric data. However, when molar concentrations were used instead, their predictive capacity raised and R^2 increased from 0.9 to 0.98 in the case of RON and MON, and from 0.85 to 0.9 in the case of S. This effect was a recurrent trend among most algorithms but it manifested with stronger intensity in linear models, which suggests that molar concentrations may be more appropriate to describe RON and MON behavior since auto-ignition reaction rates are also a function of the partial pressure of the gases in the mixture.

The use of more sophisticated tools showed an improvement with respect to traditional ones. Models built using Support Vector Regression estimators with an integrated non-linear kernel function showed the best performance for the three properties. In the case of RON and MON the SVR algorithm trained with volumetric data made the most accurate predictions, while for S models NuSVR algorithm trained on molar

data performed better. The trained models were able to capture the non-linearities within the data and predict octane values for the test set with high precision and low standard deviation. For the RON models, R^2 values over 0.99 were achieved using the testing data. In the case of MON and S, results were slightly inferior due to less data availability with R^2 equal to 0.98 and 0.94 respectively. The mean absolute error of the predicted values was always 1 unit or below. Moreover, these algorithms showed consistency and adaptability when changes were performed over the original data sets through the sensitivity analysis.

The rest of the models, based on Nearest Neighbors algorithms, Decision Trees and Random Forest resulted in satisfactory results as well, and R^2 fluctuated between 0.92 and 0.98 in most cases. The only exception to this is the r-NN algorithm, which given the used mechanism is not capable to generalize with an unevenly distributed data set like the one in this study.

6.1 Limitations and applicability

Despite the accuracy achieved by the models investigated in this study, the analysis of the results revealed some limitations mainly caused by the characteristics of the original data set. The selected species do not have an even representation in the database, which leads to higher error for those regions of the data spectrum with lower density of samples. The same logic applies to different RON, MON and S ranges. For those segments where the algorithms had few examples available, the learning rate was lower and inaccuracy became more likely during the prediction phase. However, the identification of these shortcomings can help to define strategies for future work and they should be taken into consideration to improve the database and strengthen the methodology.

An additional constrain of these models is directly linked to the type of input they are designed for, that is, neat molecules. In this regard, they are not able to represent the effect of blending different gasoline fractions in a direct way, which might result more interesting from the operational point of view of a refinery. Nevertheless, simplified approaches like the one presented in this thesis can still find application as tools to perform preliminary assessment of new gasoline blends in order to meet fuel standards in an efficient and economical way.

6.2 Future recommendations

In view of the results, increasing the number of samples in the database could potentially improve the predictions. Moreover, doing this through in-house experiments would allow to target specific blend compositions and increase the consistency and usability of the database, and consequently of the models. Additionally, considering new molecular species would lead towards a closer representation of gasoline fuels. This would facilitate the inclusion of this type of predictive models in real industrial applications, eventually enabling the creation of tailor-made models for specific process units in a refinery.

In later stages, performing more extensive investigations regarding related research questions such as how different gasoline fractions interact with each other or how various molecular characteristics like chain length or branching index shape blending behavior would be beneficial. Together with the methodology already established during the completion of this work, it will enable to develop more sophisticated instruments and facilitate the deployment of more sustainable fuels for transportation.

References

[1] IEA, "Global Energy & CO2 Status Report 2018," Tech. Rep., Mar. 2019.

[2] ——, "Statistics | World - Total Primary Energy Supply (TPES) by source (table)," https://www.iea.org/statistics/?country=WORLD&year=2016&category=Energy%20supply&indicator=TPESbySource&mode=table&dataTable=BALANCES.

[3] ——, "CO2 Emissions from Fuel Combustion 2018 Highlights," Tech. Rep., 2018.

[4] "Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R. K. Pachauri and L. A. Meyer (eds.)]," IPCC, Geneva, Switzerland, Tech. Rep., 2014.

[5] NASA, "Carbon Dioxide Concentration | NASA Global Climate Change," https://climate.nasa.gov/vital-signs/carbon-dioxide.

[6] ——, "Global Surface Temperature | NASA Global Climate Change," https://climate.nasa.gov/vital-signs/global-temperature.

[7] UNFCCC, "What is the Kyoto Protocol?" https://unfccc.int/kyoto_protocol.

[8] ——, "The Paris Agreement," https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement.

[9] European Commission, "Paris Agreement," https://ec.europa.eu/clima/policies/international/negotiations/paris_en, Nov. 2016.

[10] ——, "2030 climate & energy framework," https://ec.europa.eu/clima/policies/strategies/2030_en, Nov. 2016.

[11] ——, "White paper 2011 - Roadmap to a Single European Transport Area," https://ec.europa.eu/transport/themes/strategies/2011_white_paper_en, Sep. 2016.

[12] European Environment Agency, "Electric vehicles as a proportion of the total fleet," https://www.eea.europa.eu/data-and-maps/indicators/proportion-of-vehicle-fleet-meeting-4/assessment-2.

[13] S. Moustakidis, "Renewable Energy – Recast to 2030 (RED II)," https://ec.europa.eu/jrc/en/jec/renewable-energy-recast-2030-red-ii, Dec. 2018.

[14] H. H. Schobert, *Chemistry of Fossil Fuels and Biofuels*, ser. Cambridge Series in Chemical Engineering. Cambridge University Press, 2013.

[15] C. R. Ferguson, *Internal Combustion Engines : Applied Thermosciences*, third edition ed. Wiley, 2016.

[16] P. Doble, M. Sandercock, E. Du Pasquier, P. Petocz, C. Roux, and M. Dawson, "Classification of premium and regular gasoline by gas chromatography/mass spectrometry, principal component analysis and artificial neural networks," *Forensic Science International*, vol. 132, no. 1, pp. 26–39, Mar. 2003.

[17] S. M. Sarathy, A. Farooq, and G. T. Kalghatgi, "Recent progress in gasoline surrogate fuels," *Progress in Energy and Combustion Science*, vol. 65, pp. 67–108, Mar. 2018.

[18] CEN, "Automotive fuels - Unleaded petrol - Requirements and test methods," May 2008.

[19] "Diesel demand: Still growing globally despite Dieselgate | McKinsey & Company," https://www.mckinsey.com/industries/oil-and-gas/our-insights/petroleum-blog/diesel-demand-still-growing-globally-despite-dieselgate.

[20] FuelsEurope, "Statistical Report 2018," Tech. Rep., 2018.

[21] US Energy Information Administration, "Refining Crude Oil - Energy Explained, Your Guide To Understanding Energy - Energy Information Administration," https://www.eia.gov/energyexplained/index.php?page=oil_refining.

[22] American Institute of Chemical Engineering, "An Oil Refinery Walk-Through," Tech. Rep., May 2014.

[23] EFOA, "The European Fuel Oxygenates Association," http://www.efoa.eu, Aug. 2011.

[24] M. K. Balki, C. Sayin, and M. Canakci, "The effect of different alcohol fuels on the performance, emission and combustion characteristics of a gasoline engine," *Fuel*, vol. 115, pp. 901–906, Jan. 2014.

[25] E. Christensen, J. Yanowitz, M. Ratcliff, and R. L. McCormick, "Renewable Oxygenate Blending Effects on Gasoline Properties," *Energy & Fuels*, vol. 25, no. 10, pp. 4723–4733, Oct. 2011.

[26] O. US EPA, "Gasoline Winter Oxygenates," https://www.epa.gov/gasoline-standards/gasoline-winter-oxygenates, Aug. 2015.

[27] P. Iodice, G. Langella, and A. Amoresano, "Ethanol in gasoline fuel blends: Effect on fuel consumption and engine out emissions of SI engines in cold operating conditions," *Applied Thermal Engineering*, vol. 130, pp. 1081–1089, Feb. 2018.

[28] "Gasoline engine," https://www.britannica.com/technology/gasoline-engine.

[29] "Directive 2009/30/EC of the European Parliament and of the Council of 23 April 2009 amending Directive 98/70/EC as regards the specification of petrol, diesel and gas-oil and introducing a mechanism to monitor and reduce greenhouse gas emissions and amending Council Directive 1999/32/EC as regards the specification of fuel used by inland waterway vessels and repealing Directive 93/12/EEC (Text with EEA relevance)," http://data.europa.eu/eli/dir/2009/30/oj/eng, Jun. 2009.

[30] EEA, "Fuel quality in the EU in 2016," Publication 24/2017, 2018.

[31] S. J. Rand, *Significance of Tests for Petroleum Products*, 8th ed. West Conshohocken, Pa: ASTM International, 2011, oCLC: 809129241.

[32] M. R. Riazi, T. A. Albahri, and A. H. Alqattan, "Prediction of Reid Vapor Pressure of Petroleum Fuels," *Petroleum Science and Technology*, vol. 23, no. 1, pp. 75–86, Dec. 2005.

[33] T. J. Bruno, A. Wolk, and A. Naydich, "Composition-Explicit Distillation Curves for Mixtures of Gasoline with Four-Carbon Alcohols (Butanols)," *Energy & Fuels*, vol. 23, no. 4, pp. 2295–2306, Apr. 2009.

[34] R. Stradling, J. Antunez, A. Bellier, C. Bomholt, N. Elliott, P. Gomez-Acebo, H. Hovius, A. Joedicke, U. Kiiski, M. Santiago, H. P. Saavedra, W. Mirabella, P. Scott, K. Skaardalsmo, S. McArragher, D. J. Rickeard, P. J. Zemroch, K. D. Rose, L. Kennedy, J. Edwards, and P. Stones, "Gasoline volatility and vehicle performance," CONCAWE, Tech. Rep. 2/12, Feb. 2012.

[35] J. P. Szybist and D. A. Splitter, "Understanding chemistry-specific fuel differences at a constant RON in a boosted SI engine," *Fuel*, vol. 217, pp. 370–381, Apr. 2018.

[36] W. R. Leppard, "The Chemical Origin of Fuel Octane Sensitivity," SAE International, Warrendale, PA, SAE Technical Paper 902137, Oct. 1990.

[37] M. Mehl, T. Faravelli, F. Giavazzi, E. Ranzi, P. Scorletti, A. Tardani, and D. Terna, "Detailed Chemistry Promotes Understanding of Octane Numbers and Gasoline Sensitivity," *Energy & Fuels*, vol. 20, no. 6, pp. 2391–2398, Nov. 2006.

[38] A. D. B. Yates, A. Swarts, and C. L. Viljoen, "Correlating Auto-Ignition Delays And Knock-Limited Spark-Advance Data For Different Types Of Fuel," SAE International, Warrendale, PA, SAE Technical Paper 2005-01-2083, May 2005.

[39] G. T. Kalghatgi, "Fuel Anti-Knock Quality - Part I. Engine Studies," SAE International, Warrendale, PA, SAE Technical Paper 2001-01-3584, Sep. 2001.

[40] P. Ghosh, K. Hickey, and S. B. Jaffe, "Development of a Detailed Gasoline Composition-Based Octane Model."

[41] T. J. Truex, "Interaction of Sulfur with Automotive Catalysts and the Impact on Vehicle Emissions-A Review," *SAE Transactions*, vol. 108, pp. 1192–1206, 1999.

[42] A. Pandey, C. Larroche, S. C. Ricke, C.-G. Dussap, and E. Gnansounou, *Biofuels: Alternative Feedstocks and Conversion Processes.* Burlington: Elsevier Science, 2011, oCLC: 742333629.

[43] V. Babu, A. Thapliyal, and G. K. Patel, *Biofuels Production.* Somerset, UNITED STATES: John Wiley & Sons, Incorporated, 2013.

[44] K. Dutta, A. Daverey, and J.-G. Lin, "Evolution retrospective for alternative fuels: First to fourth generation," *Renewable Energy*, vol. 69, pp. 114–122, Sep. 2014.

[45] J. A. Quintero, M. I. Montoya, O. J. Sánchez, O. H. Giraldo, and C. A. Cardona, "Fuel ethanol production from sugarcane and corn: Comparative analysis for a Colombian case," *Energy*, vol. 33, no. 3, pp. 385–399, Mar. 2008.

[46] "Assesment of the impact of ethanol content in gasoline on fuel consumption, includign a literature review up to 2006," CONCAWE, Brussels, Tech. Rep. 13/13, Dec. 2013.

[47] US Department of Energy, "Alternative Fuels Data Center: Flexible Fuel Vehicles," https://afdc.energy.gov/vehicles/flexible_fuel.html.

[48] Q. Nguyen, J. Bowter, J. Howe, S. Bratkovich, H. Groot, E. Pepke, and K. Fernholz, "Global production of second generation biofuels: Trensd and influences," Dovetail Partners, INC., Tech. Rep., 2017.

[49] "From 1st to 2nd generation biofuel technologies. An overview of current industry and RD&D activities," OECD/IEA, Tech. Rep., 2008.

[50] R. Sands and S. Malcolm, "Dedicating Agricultural Land to Energy Crops Would Shift Land Use," https://www.ers.usda.gov/amber-waves/2017/april/dedicating-agricultural-land-to-energy-crops-would-shift-land-use/.

[51] C. Baranzelli, C. Perpiña Castillo, A. Lopes Barbosa, F. Batista e Silva, C. Jacobs-Crisioni, and C. Lavalle, "Land allocation and suitability analysis for the production of food, feed and energy crops in the period 2010 - 2050," European Commission, Text JRC98567, 2015.

[52] P. Das and P. Tiwari, "The effect of slow pyrolysis on the conversion of packaging waste plastics (PE and PP) into fuel," *Waste Management*, vol. 79, pp. 615–624, Sep. 2018.

[53] P. Tandon and Q. Jin, "Microalgae culture enhancement through key microbial approaches," *Renewable and Sustainable Energy Reviews*, vol. 80, pp. 1089–1099, Dec. 2017.

[54] W. Li, H. Wang, X. Jiang, J. Zhu, Z. Liu, X. Guo, and C. Song, "A short review of recent advances in CO2 hydrogenation to hydrocarbons over heterogeneous catalysts," *RSC Advances*, vol. 8, no. 14, pp. 7651–7669, Feb. 2018.

[55] M. J. Biddy, R. Davis, D. Humbird, L. Tao, N. Dowe, M. T. Guarnieri, J. G. Linger, E. M. Karp, D. Salvachúa, D. R. Vardon, and G. T. Beckham, "The Techno-Economic Basis for Coproduct Manufacturing To Enable Hydrocarbon Fuel Production from Lignocellulosic Biomass," *ACS Sustainable Chemistry & Engineering*, vol. 4, no. 6, pp. 3196–3211, Jun. 2016.

[56] R. Hilten, R. Speir, J. Kastner, and K. C. Das, "Production of aromatic green gasoline additives via catalytic pyrolysis of acidulated peanut oil soap stock," *Bioresource Technology*, vol. 102, no. 17, pp. 8288–8294, Sep. 2011.

[57] Y. Kar, "Pyrolysis of waste pomegranate peels for bio-oil production," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 40, no. 23, pp. 2812–2821, 2018.

[58] G. Duman, C. Okutucu, S. Ucar, R. Stahl, and J. Yanik, "The slow and fast pyrolysis of cherry seed," *Bioresource Technology*, vol. 102, no. 2, pp. 1869–1878, Jan. 2011.

[59] S. A. Hanafi, M. S. Elmelawy, N. H. Shalaby, H. A. El-Syed, G. Eshaq, and M. S. Mostafa, "Hydrocracking of waste chicken fat as a cost effective feedstock for renewable fuel production: A kinetic study," *Egyptian Journal of Petroleum*, vol. 25, no. 4, pp. 531–537, Dec. 2016.

[60] V. R. Wiggers, A. Wisniewski, L. A. S. Madureira, A. A. C. Barros, and H. F. Meier, "Biofuels from waste fish oil pyrolysis: Continuous production in a pilot plant," *Fuel*, vol. 88, no. 11, pp. 2135–2141, Nov. 2009.

[61] J. M. Jarvis, K. O. Albrecht, J. M. Billing, A. J. Schmidt, R. T. Hallen, and T. M. Schaub, "Assessment of Hydrotreatment for Hydrothermal Liquefaction Biocrudes from Sewage Sludge, Microalgae, and Pine Feedstocks," *Energy & Fuels*, vol. 32, no. 8, pp. 8483–8493, Aug. 2018.

[62] R. C. Brown and C. Stevens, *Thermochemical Processing of Biomass: Conversion into Fuels, Chemicals and Power*. Hoboken, UNITED KINGDOM: John Wiley & Sons, Incorporated, 2011.

[63] D. Castello, M. S. Haider, and L. A. Rosendahl, "Catalytic upgrading of hydrothermal liquefaction biocrudes: Different challenges for different feedstocks," *Renewable Energy*, vol. 141, pp. 420–430, Oct. 2019.

[64] C. Liu, H. Wang, A. M. Karim, J. Sun, and Y. Wang, "Catalytic fast pyrolysis of lignocellulosic biomass," *Chemical Society Reviews*, vol. 43, no. 22, pp. 7594–7623, Oct. 2014.

[65] V. Chiodo, G. Zafarana, S. Maisano, S. Freni, and F. Urbani, "Pyrolysis of different biomass: Direct comparison among Posidonia Oceanica, Lacustrine Alga and White-Pine," *Fuel*, vol. 164, pp. 220–227, Jan. 2016.

[66] J. Bengtsson and S.-L. Nonås, "Refinery planning and scheduling - An overview," Institute for research in economics and business administration, Bergen, Tech. Rep. 29/08, 2009.

[67] S. Jiang, "Optimisation of diesel and gasoline blending operations," Ph.D. Dissertation, Faculty of Engineering and Physical Sciences, University of Manchester, 2016.

[68] W. G. Lovell, J. M. Campbell, and T. A. Boyd, "Detonation Characteristics of Some Aliphatic Olefin Hydrocarbons," *Industrial & Engineering Chemistry*, vol. 23, no. 5, pp. 555–558, May 1931.

[69] M. D. Boot, M. Tian, E. J. M. Hensen, and S. Mani Sarathy, "Impact of fuel molecular structure on auto-ignition behavior – Design rules for future high performance gasolines," *Progress in Energy and Combustion Science*, vol. 60, pp. 1–25, May 2017.

[70] E. Monroe, J. Gladden, K. O. Albrecht, J. T. Bays, R. McCormick, R. W. Davis, and A. George, "Discovery of novel octane hyperboosting phenomenon in prenol biofuel/gasoline blends," *Fuel*, vol. 239, pp. 1143–1148, Mar. 2019.

[71] T. A. Albahri, "Structural Group Contribution Method for Predicting the Octane Number of Pure Hydrocarbon Liquids." *Industrial & Engineering Chemistry Research*, vol. 43, no. 24, pp. 7964–7964, Nov. 2004.

[72] A. L. Lapidus, E. A. Smolenskii, V. M. Bavykin, T. N. Myshenkova, and L. T. Kondrat'ev, "Models for the calculation and prediction of the octane and cetane numbers of individual hydrocarbons," *Petroleum Chemistry*, vol. 48, no. 4, pp. 277–286, Jul. 2008.

[73] S. R. Daly, K. E. Niemeyer, W. J. Cannella, and C. L. Hagen, "Predicting fuel research octane number using Fourier-transform infrared absorption spectra of neat hydrocarbons," *Fuel*, vol. 183, pp. 359–365, Nov. 2016.

[74] T. M. Foong, K. J. Morganti, M. J. Brear, G. da Silva, Y. Yang, and F. L. Dryer, "The octane numbers of ethanol blended with gasoline and its surrogates," *Fuel*, vol. 115, pp. 727–739, Jan. 2014.

[75] J. J. Kelly, C. H. Barlow, T. M. Jinguji, and J. B. Callis, "Prediction of gasoline octane numbers from near-infrared spectral features in the range 660-1215 nm," *Analytical Chemistry*, vol. 61, no. 4, pp. 313–320, Feb. 1989.

[76] D. Özdemir, "Determination of Octane Number of Gasoline Using Near Infrared Spectroscopy and Genetic Multivariate Calibration Methods," *Petroleum Science and Technology*, vol. 23, no. 9-10, pp. 1139–1152, Sep. 2005.

[77] A. G. Abdul Jameel, V. Van Oudenhoven, A.-H. Emwas, and S. M. Sarathy, "Predicting Octane Number Using Nuclear Magnetic Resonance Spectroscopy and Artificial Neural Networks," *Energy & Fuels*, vol. 32, no. 5, pp. 6309–6329, May 2018.

[78] G. Protić-Lovasić, N. Jambrec, D. Deur-Siftar, and M. V. Prostenik, "Determination of catalytic reformed gasoline octane number by high resolution gas chromatography," *Fuel*, vol. 69, no. 4, pp. 525–528, Apr. 1990.

[79] S. J. Alexandrovna and Duong Chi Tuyen, "Development of a detailed model for calculating the octane numbers of gasoline blends," in *International Forum on Strategic Technology 2010*, Oct. 2010, pp. 430–432.

[80] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, pp. 210–229, 1959.

[81] D. Bzdok, N. Altman, and M. Krzywinski, "Statistics versus machine learning," *Nature Methods*, vol. 15, p. 233, Apr. 2018.

[82] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 1st ed. O'Reilly Media, Inc, 2017.

[83] L. Massaron and A. Boschetti, *Regression Analysis with Python*, 1st ed. Packt Publishing, 2016.

[84] A. Müller, *Introduction to Machine Learning with Python*, 1st ed. O'Reilly Media, Inc, 2016.

[85] "Supervised learning — scikit-learn 0.21.3 documentation," https://scikit-learn.org/stable/supervised_learning.html#supervised-learning.

[86] G. Drakos, "Support Vector Machine vs Logistic Regression," https://towardsdatascience.com/support-vector-machine-vs-logistic-regression-94cc2975433f, Oct. 2018.

[87] M. Mohammed, *Machine Learning*, 1st ed. CRC Press, 2016.

[88] D. Wilimitis, "The Kernel Trick in Support Vector Classification," https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f, Feb. 2019.

[89] Z. Liu, L. Zhang, A. Elkamel, D. Liang, S. Zhao, C. Xu, S. Y. Ivanov, and A. K. Ray, "Multiobjective Feature Selection Approach to Quantitative Structure Property Relationship Models for Predicting the Octane Number of Compounds Found in Gasoline," *Energy & Fuels*, vol. 31, no. 6, pp. 5828–5839, Jun. 2017.

[90] A. R, "APPLYING RANDOM FOREST (CLASSIFICATION) — MACHINE LEARNING ALGORITHM FROM SCRATCH WITH REAL DATASETS," https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c57, Jul. 2018.

[91] S. Lee, H. Choi, K. Cha, and H. Chung, "Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha," *Microchemical Journal*, vol. 110, pp. 739–748, Sep. 2013.

[92] "Nervous System | The Partnership in Education," https://www.thepartnershipineducation.com/resources/nervous-system.

[93] T. A. Albahri, "Specific Gravity, RVP, Octane Number, and Saturates, Olefins, and Aromatics Fractional Composition of Gasoline and Petroleum Fractions by Neural Network Algorithms," *Petroleum Science and Technology*, vol. 32, no. 10, pp. 1219–1226, May 2014.

[94] M. Ferreiro-González, J. Ayuso, J. A. Álvarez, M. Palma, and C. G. Barroso, "New Headspace-Mass Spectrometry Method for the Discrimination of Commercial Gasoline Samples with Different Research Octane Numbers," *Energy & Fuels*, vol. 28, no. 10, pp. 6249–6254, Oct. 2014.

[95] "DBSCAN," *Wikipedia*, Sep. 2019, page Version ID: 916286433.

[96] "DBSCAN: Density-based clustering for discovering clusters in large datasets with noise - Unsupervised Machine Learning - Easy Guides - Wiki - STHDA," http://www.sthda.com/english/wiki/wiki.php?id_contents=7940.

[97] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2015, pp. 1200–1205.

[98] J. Badra, A. S. AlRamadan, and S. M. Sarathy, "Optimization of the octane response of gasoline/ethanol blends," *Applied Energy*, vol. 203, pp. 778–793, Oct. 2017.

[99] E. Hu, J. Ku, G. Yin, C. Li, X. Lu, and Z. Huang, "Laminar Flame Characteristics and Kinetic Modeling Study of Ethyl Tertiary Butyl Ether Compared with Methyl Tertiary Butyl Ether, Ethanol, iso-Octane, and Gasoline," *Energy & Fuels*, vol. 32, no. 3, pp. 3935–3949, Mar. 2018.

[100] W. M. Haynes and D. R. Lide, Eds., *CRC Handbook of Chemistry and Physics: A Ready-Reference Book of Chemical and Physical Data*, 96th ed. Boca Raton, Fla.: CRC Press, 2015, oCLC: 927175199.

[101] N. Ré-Poppi, F. F. P. Almeida, C. A. L. Cardoso, J. L. Raposo, L. H. Viana, T. Q. Silva, J. L. C. Souza, and V. S. Ferreira, "Screening analysis of type C Brazilian gasoline by gas chromatography – Flame ionization detector," *Fuel*, vol. 88, no. 3, pp. 418–423, Mar. 2009.

[102] G. Kalghatgi, H. Babiker, and J. Badra, "A Simple Method to Predict Knock Using Toluene, N-Heptane and Iso-Octane Blends (TPRF) as Gasoline Surrogates," *SAE International Journal of Engines*, vol. 8, no. 2, pp. 505–519, Apr. 2015.

[103] Developer Economics, "What is the best programming language for Machine Learning?" https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7, Jan. 2019.

[104] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Oct. 2011.

[105] D. Shulga, "5 Reasons why you should use Cross-Validation in your Data Science Projects," https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79, Sep. 2018.

[106] A. Ng, "What data scientists should know about deep learning," 2015.

[107] J. E. Anderson, U. Kramer, S. A. Mueller, and T. J. Wallington, "Octane Numbers of Ethanol- and Methanol-Gasoline Blends Estimated from Molar Concentrations," *Energy & Fuels*, vol. 24, no. 12, pp. 6576–6585, Dec. 2010.

[108] J. W. G. Turner, R. J. Pearson, A. Bell, S. de Goede, and C. Woolard, "Iso-Stoichiometric Ternary Blends of Gasoline, Ethanol and Methanol: Investigations into Exhaust Emissions, Blend Properties and Octane Numbers," *SAE International Journal of Fuels and Lubricants*, vol. 5, no. 3, pp. 945–967, Sep. 2012.

[109] K. L, Ž. A, B. R, S. A, N. G, K. Ž, D. N, J. R, and J. G, "Experimental determination of distillation curves of alcohols/gasoline blends as bio-fuel for SI engines," *Machines. Technologies. Materials.*, vol. 9, no. 8, pp. 18–21, 2015.

[110] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the Surprising Behavior of Distance Metrics in High Dimensional Space," in *Database Theory — ICDT 2001*, ser. Lecture Notes in Computer Science, J. Van den Bussche and V. Vianu, Eds. Springer Berlin Heidelberg, 2001, pp. 420–434.

[111] J. A. Badra, N. Bokhumseen, N. Mulla, S. M. Sarathy, A. Farooq, G. Kalghatgi, and P. Gaillard, "A methodology to relate octane numbers of binary and ternary n-heptane, iso-octane and toluene mixtures with simulated ignition delay times," *Fuel*, vol. 160, pp. 458–469, Nov. 2015.

[112] S. M. Sarathy, G. Kukkadapu, M. Mehl, T. Javed, A. Ahmed, N. Naser, A. Tekawade, G. Kosiba, M. AlAbbad, E. Singh, S. Park, M. A. Rashidi, S. H. Chung, W. L. Roberts, M. A. Oehlschlaeger, C.-J. Sung, and A. Farooq, "Compositional effects on the ignition of FACE gasolines," *Combustion and Flame*, vol. 169, pp. 171–193, Jul. 2016.

[113] B. Li and Y. Jiang, "Chemical Kinetic Model of a Multicomponent Gasoline Surrogate with Cross Reactions," *Energy & Fuels*, vol. 32, no. 9, pp. 9859–9871, Sep. 2018.

[114] E. Singh, J. Badra, M. Mehl, and S. M. Sarathy, "Chemical Kinetic Insights into the Octane Number and Octane Sensitivity of Gasoline Surrogate Mixtures," *Energy & Fuels*, vol. 31, no. 2, pp. 1945–1960, Feb. 2017.

[115] B. Wolk, I. Ekoto, and W. Northrop, "Investigation of Fuel Effects on In-Cylinder Reforming Chemistry Using Gas Chromatography," *SAE International Journal of Engines*, vol. 9, no. 2, pp. 964–978, Apr. 2016.

[116] P. L. Perez and A. L. Boehman, "Experimental Investigation of the Autoignition Behavior of Surrogate Gasoline Fuels in a Constant-Volume Combustion Bomb Apparatus and Its Relevance to HCCI Combustion," *Energy & Fuels*, vol. 26, no. 10, pp. 6106–6117, Oct. 2012.

[117] D. M. Cameron, "Autoignition Studies of Gasoline Surrogate Fuels in the Advanced Fuel Ignition Delay Analyzer," p. 109.

[118] T. Ogura, Y. Sakai, A. Miyoshi, M. Koshi, and P. Dagaut, "Modeling of the Oxidation of Primary Reference Fuel in the Presence of Oxygenated Octane Improvers: Ethyl Tert-Butyl Ether and Ethanol," *Energy & Fuels*, vol. 21, no. 6, pp. 3233–3239, Nov. 2007.

[119] L. R. Cancino, M. Fikri, A. A. M. Oliveira, and C. Schulz, "Autoignition of gasoline surrogate mixtures at intermediate temperatures and high pressures: Experimental and numerical approaches," *Proceedings of the Combustion Institute*, vol. 32, no. 1, pp. 501–508, Jan. 2009.

[120] N. Morgan, A. Smallbone, A. Bhave, M. Kraft, R. Cracknell, and G. Kalghatgi, "Mapping surrogate gasoline compositions into RON/MON space," *Combustion and Flame*, vol. 157, no. 6, pp. 1122–1131, Jun. 2010.

Appendix 1. Database

id	#	P	I	O	N	A	Ox	Ether	RON	MON	S	Ref
1	1	100.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0	0	[31]
2	1	0.00	100.00	0.00	0.00	0.00	0.00	0.00	100	100	0	[31]
3	1	0.00	0.00	100.00	0.00	0.00	0.00	0.00	73.6	64.5	9.1	[98]
4	1	0.00	0.00	0.00	100.00	0.00	0.00	0.00	100	84.9	15.1	[40]
5	1	0.00	0.00	0.00	0.00	100.00	0.00	0.00	118	103.5	14.5	[40]
6	1	0.00	0.00	0.00	0.00	0.00	100.00	0.00	108	92.9	15.1	[40]
7	1	0.00	0.00	0.00	0.00	0.00	0.00	100.00	117	101	16	[99]
8	2	5.00	95.00	0.00	0.00	0.00	0.00	0.00	95	95	0	[31]
9	2	10.00	90.00	0.00	0.00	0.00	0.00	0.00	90	90	0	[31]
10	2	15.00	85.00	0.00	0.00	0.00	0.00	0.00	85	85	0	[31]
11	2	20.00	80.00	0.00	0.00	0.00	0.00	0.00	80	80	0	[31]
12	2	25.00	75.00	0.00	0.00	0.00	0.00	0.00	75	75	0	[31]
13	2	30.00	70.00	0.00	0.00	0.00	0.00	0.00	70	70	0	[31]
14	2	35.00	65.00	0.00	0.00	0.00	0.00	0.00	65	65	0	[31]
15	2	40.00	60.00	0.00	0.00	0.00	0.00	0.00	60	60	0	[31]
16	2	45.00	55.00	0.00	0.00	0.00	0.00	0.00	55	55	0	[31]
17	2	50.00	50.00	0.00	0.00	0.00	0.00	0.00	50	50	0	[31]
18	2	55.00	45.00	0.00	0.00	0.00	0.00	0.00	45	45	0	[31]
19	2	60.00	40.00	0.00	0.00	0.00	0.00	0.00	40	40	0	[31]
20	2	65.00	35.00	0.00	0.00	0.00	0.00	0.00	35	35	0	[31]
21	2	70.00	30.00	0.00	0.00	0.00	0.00	0.00	30	30	0	[31]
22	2	75.00	25.00	0.00	0.00	0.00	0.00	0.00	25	25	0	[31]
23	2	80.00	20.00	0.00	0.00	0.00	0.00	0.00	20	20	0	[31]
24	2	85.00	15.00	0.00	0.00	0.00	0.00	0.00	15	15	0	[31]
25	2	90.00	10.00	0.00	0.00	0.00	0.00	0.00	10	10	0	[31]
26	2	95.00	5.00	0.00	0.00	0.00	0.00	0.00	5	5	0	[31]
27	3	72.00	18.00	0.00	0.00	10.00	0.00	0.00	32	NaN	NaN	[102]
28	3	64.00	16.00	0.00	0.00	20.00	0.00	0.00	42	NaN	NaN	[102]
29	3	56.00	14.00	0.00	0.00	30.00	0.00	0.00	53.2	48	5.2	[102]
30	3	48.00	12.00	0.00	0.00	40.00	0.00	0.00	63.7	58	5.7	[102]
31	3	40.00	10.00	0.00	0.00	50.00	0.00	0.00	75.5	68	7.5	[102]
32	3	21.00	5.00	0.00	0.00	74.00	0.00	0.00	96.9	85.2	11.7	[102]
33	3	16.00	10.00	0.00	0.00	74.00	0.00	0.00	99.8	88.7	11.1	[102]
34	3	54.00	36.00	0.00	0.00	10.00	0.00	0.00	48	NaN	NaN	[102]
35	3	48.00	32.00	0.00	0.00	20.00	0.00	0.00	58	NaN	NaN	[102]
36	3	36.00	24.00	0.00	0.00	40.00	0.00	0.00	75.1	68	7.1	[102]
37	3	30.00	20.00	0.00	0.00	50.00	0.00	0.00	83.8	76.2	7.6	[102]
38	3	42.00	28.00	0.00	0.00	30.00	0.00	0.00	66.1	61	5.1	[102]
39	3	36.00	54.00	0.00	0.00	10.00	0.00	0.00	66	64.4	1.6	[102]
40	3	32.00	48.00	0.00	0.00	20.00	0.00	0.00	73.6	70	3.6	[102]
41	3	24.00	36.00	0.00	0.00	40.00	0.00	0.00	86.2	79.6	6.6	[102]
42	3	20.00	30.00	0.00	0.00	50.00	0.00	0.00	92.1	82.9	9.2	[102]
43	3	28.00	42.00	0.00	0.00	30.00	0.00	0.00	79	74	5	[102]
44	3	16.00	64.00	0.00	0.00	20.00	0.00	0.00	89.1	85.6	3.5	[102]
45	3	14.00	56.00	0.00	0.00	30.00	0.00	0.00	92.8	86.9	5.9	[102]
46	3	12.00	48.00	0.00	0.00	40.00	0.00	0.00	96.7	88.7	8	[102]
47	3	10.00	40.00	0.00	0.00	50.00	0.00	0.00	99.8	90.9	8.9	[102]
48	3	18.00	72.00	0.00	0.00	10.00	0.00	0.00	84.5	82	2.5	[102]
49	2	42.00	0.00	0.00	0.00	58.00	0.00	0.00	75.6	66.9	8.7	[102]
50	2	34.00	0.00	0.00	0.00	66.00	0.00	0.00	85.2	74.8	10.4	[102]
51	2	30.00	0.00	0.00	0.00	70.00	0.00	0.00	89.3	78.2	11.1	[102]
52	2	26.00	0.00	0.00	0.00	74.00	0.00	0.00	93.4	81.5	11.9	[102]
53	3	30.00	10.00	0.00	0.00	60.00	0.00	0.00	85.3	75.2	10.1	[102]
54	3	20.00	20.00	0.00	0.00	60.00	0.00	0.00	95	83.7	11.3	[102]
55	3	64.00	17.00	0.00	0.00	19.00	0.00	0.00	39	37	2	[102]
56	2	50.00	0.00	0.00	0.00	50.00	0.00	0.00	65.9	57.7	8.2	[102]
57	3	33.33	33.33	0.00	0.00	33.33	0.00	0.00	76.2	70.9	5.3	[102]
58	3	17.00	67.00	0.00	0.00	16.00	0.00	0.00	87	84	3	[102]
59	3	16.67	16.67	0.00	0.00	66.67	0.00	0.00	98	87.4	10.6	[102]
60	2	27.00	0.00	0.00	0.00	73.00	0.00	0.00	92.3	80.7	11.6	[102]
61	3	10.00	72.00	0.00	0.00	18.00	0.00	0.00	93.7	90.3	3.4	[102]
62	3	16.83	43.56	0.00	0.00	39.60	0.00	0.00	93	85.8	7.2	[102]
63	3	14.85	51.49	0.00	0.00	33.66	0.00	0.00	93	86.7	6.3	[102]
64	2	21.00	0.00	0.00	0.00	79.00	0.00	0.00	97.7	86.2	11.5	[102]
65	3	10.00	65.00	0.00	0.00	25.00	0.00	0.00	95.2	90.5	4.7	[102]
66	3	15.00	35.00	0.00	0.00	50.00	0.00	0.00	96.3	87.3	9	[102]

Continued on next page

id	#	P	I	O	N	A	Ox	Ether	RON	MON	S	Ref
67	3	17.00	69.00	0.00	0.00	14.00	0.00	0.00	86.6	84.2	2.4	[102]
68	3	16.00	74.00	0.00	0.00	10.00	0.00	0.00	85.7	84.6	1.1	[102]
69	3	13.86	42.57	0.00	0.00	43.56	0.00	0.00	96.3	88.3	8	[102]
70	2	0.00	90.00	0.00	0.00	10.00	0.00	0.00	102	NaN	NaN	[102]
71	2	0.00	80.00	0.00	0.00	20.00	0.00	0.00	104.1	NaN	NaN	[102]
72	2	0.00	70.00	0.00	0.00	30.00	0.00	0.00	105.6	NaN	NaN	[102]
73	2	0.00	60.00	0.00	0.00	40.00	0.00	0.00	107.7	NaN	NaN	[102]
74	2	0.00	50.00	0.00	0.00	50.00	0.00	0.00	108.2	100.3	7.9	[102]
75	3	10.00	30.00	0.00	0.00	60.00	0.00	0.00	101.6	91.2	10.4	[102]
76	2	0.00	40.00	0.00	0.00	60.00	0.00	0.00	110	100.4	9.6	[102]
77	3	16.00	4.00	0.00	0.00	80.00	0.00	0.00	101	90.5	10.5	[102]
78	3	12.00	8.00	0.00	0.00	80.00	0.00	0.00	103.1	94	9.1	[102]
79	3	8.00	12.00	0.00	0.00	80.00	0.00	0.00	105.4	97.5	7.9	[102]
80	3	2.00	18.00	0.00	0.00	80.00	0.00	0.00	108.5	101	7.5	[102]
81	2	0.00	20.00	0.00	0.00	80.00	0.00	0.00	112.6	102.8	9.8	[102]
82	2	10.00	0.00	0.00	0.00	90.00	0.00	0.00	106	100	6	[102]
83	3	8.00	2.00	0.00	0.00	90.00	0.00	0.00	108	101.4	6.6	[102]
84	3	6.00	4.00	0.00	0.00	90.00	0.00	0.00	109.5	102.4	7.1	[102]
85	3	4.00	6.00	0.00	0.00	90.00	0.00	0.00	111.8	104.4	7.4	[102]
86	3	11.00	15.00	0.00	0.00	74.00	0.00	0.00	103.3	92.6	10.7	[102]
87	3	6.00	20.00	0.00	0.00	74.00	0.00	0.00	107.6	96.6	11	[102]
88	3	16.67	16.67	0.00	0.00	66.67	0.00	0.00	110	99.3	10.7	[102]
89	2	90.00	0.00	0.00	0.00	10.00	0.00	0.00	22	NaN	NaN	[111]
90	3	54.00	36.00	0.00	0.00	10.00	0.00	0.00	48	46.7	1.3	[111]
91	2	0.00	90.00	0.00	0.00	10.00	0.00	0.00	102	99.5	2.5	[111]
92	2	80.00	0.00	0.00	0.00	20.00	0.00	0.00	28	NaN	NaN	[111]
93	3	48.00	32.00	0.00	0.00	20.00	0.00	0.00	58	53.9	4.1	[111]
94	2	0.00	80.00	0.00	0.00	20.00	0.00	0.00	104.1	98.5	5.6	[111]
95	2	70.00	0.00	0.00	0.00	30.00	0.00	0.00	38	NaN	NaN	[111]
96	2	0.00	70.00	0.00	0.00	30.00	0.00	0.00	105.6	98.3	7.3	[111]
97	2	60.00	0.00	0.00	0.00	40.00	0.00	0.00	51.4	46	5.4	[111]
98	2	0.00	60.00	0.00	0.00	40.00	0.00	0.00	107.7	97	10.7	[111]
99	2	50.00	0.00	0.00	0.00	50.00	0.00	0.00	65.9	60	5.9	[111]
100	2	40.00	0.00	0.00	0.00	60.00	0.00	0.00	77	68	9	[111]
101	2	20.00	0.00	0.00	0.00	80.00	0.00	0.00	98.6	88.5	10.1	[111]
102	3	36.00	54.00	0.00	0.00	10.00	0.00	0.00	66	64.4	1.6	[111]
103	3	42.00	28.00	0.00	0.00	30.00	0.00	0.00	66.1	61	5.1	[111]
104	2	50.00	0.00	0.00	0.00	50.00	0.00	0.00	65.9	60	5.9	[111]
105	3	12.40	72.60	0.00	0.00	15.00	0.00	0.00	91	88.4	2.6	[74]
106	3	17.00	53.20	0.00	0.00	29.80	0.00	0.00	91.3	86.1	5.2	[74]
107	3	20.30	34.70	0.00	0.00	45.00	0.00	0.00	91.1	83.5	7.6	[74]
108	2	0.00	0.00	0.00	0.00	90.00	10.00	0.00	112.8	101	11.8	[74]
109	2	0.00	0.00	0.00	0.00	80.00	20.00	0.00	110.9	97	13.9	[74]
110	2	0.00	0.00	0.00	0.00	60.00	40.00	0.00	108.6	93.3	15.3	[74]
111	2	0.00	0.00	0.00	0.00	40.00	60.00	0.00	108.1	91.9	16.2	[74]
112	2	0.00	0.00	0.00	0.00	20.00	80.00	0.00	107.9	91.1	16.8	[74]
113	3	8.10	81.90	0.00	0.00	0.00	10.00	0.00	98.7	94.3	4.4	[74]
114	3	7.20	72.80	0.00	0.00	0.00	20.00	0.00	103.8	95.3	8.5	[74]
115	3	5.40	54.60	0.00	0.00	0.00	40.00	0.00	108	94.5	13.5	[74]
116	3	3.60	36.40	0.00	0.00	0.00	60.00	0.00	108.4	93.4	15	[74]
117	3	1.80	18.20	0.00	0.00	0.00	80.00	0.00	108.4	92.2	16.2	[74]
118	4	15.30	47.88	0.00	0.00	26.82	10.00	0.00	97	89.4	7.6	[74]
119	4	13.60	42.56	0.00	0.00	23.84	20.00	0.00	101.4	91.1	10.3	[74]
120	4	10.20	31.92	0.00	0.00	17.88	40.00	0.00	106	92.1	13.9	[74]
121	4	6.80	21.28	0.00	0.00	11.92	60.00	0.00	107.1	92	15.1	[74]
122	4	3.40	10.64	0.00	0.00	5.96	80.00	0.00	107.5	91.4	16.1	[74]
123	4	11.16	65.34	0.00	0.00	13.50	10.00	0.00	97.8	91.7	6.1	[74]
124	4	9.92	58.08	0.00	0.00	12.00	20.00	0.00	102.6	93.2	9.4	[74]
125	4	7.44	43.56	0.00	0.00	9.00	40.00	0.00	107.1	93.6	13.5	[74]
126	4	4.96	29.04	0.00	0.00	6.00	60.00	0.00	107.7	92.6	15.1	[74]
127	4	2.48	14.52	0.00	0.00	3.00	80.00	0.00	107.8	91.7	16.1	[74]
128	4	18.27	31.23	0.00	0.00	40.50	10.00	0.00	96	87.2	8.8	[74]
129	4	16.24	27.76	0.00	0.00	36.00	20.00	0.00	100.2	89.1	11.1	[74]
130	4	12.18	20.82	0.00	0.00	27.00	40.00	0.00	104.6	90.9	13.7	[74]
131	4	8.12	13.88	0.00	0.00	18.00	60.00	0.00	106.3	91.2	15.1	[74]
132	4	4.06	6.94	0.00	0.00	9.00	80.00	0.00	107.1	91.1	16	[74]
133	2	70.00	0.00	0.00	0.00	0.00	30.00	0.00	54.3	NaN	NaN	[74]
134	2	60.00	0.00	0.00	0.00	0.00	40.00	0.00	69.7	64.5	5.2	[74]
135	2	50.00	0.00	0.00	0.00	0.00	50.00	0.00	83.8	NaN	NaN	[74]

Continued on next page

id	#	P	I	O	N	A	Ox	Ether	RON	MON	S	Ref
136	2	40.00	0.00	0.00	0.00	0.00	60.00	0.00	94.7	83.8	10.9	[74]
137	2	30.00	0.00	0.00	0.00	0.00	70.00	0.00	101.6	NaN	NaN	[74]
138	2	20.00	0.00	0.00	0.00	0.00	80.00	0.00	104.7	88.9	15.8	[74]
139	2	10.00	0.00	0.00	0.00	0.00	90.00	0.00	106.5	NaN	NaN	[74]
140	3	72.00	8.00	0.00	0.00	0.00	20.00	0.00	45.9	NaN	NaN	[74]
141	3	63.00	7.00	0.00	0.00	0.00	30.00	0.00	61.1	NaN	NaN	[74]
142	3	54.00	6.00	0.00	0.00	0.00	40.00	0.00	75.6	NaN	NaN	[74]
143	3	45.00	5.00	0.00	0.00	0.00	50.00	0.00	87.6	NaN	NaN	[74]
144	3	36.00	4.00	0.00	0.00	0.00	60.00	0.00	96.6	NaN	NaN	[74]
145	3	64.00	16.00	0.00	0.00	0.00	20.00	0.00	53.3	NaN	NaN	[74]
146	3	56.00	14.00	0.00	0.00	0.00	30.00	0.00	67.4	NaN	NaN	[74]
147	3	48.00	12.00	0.00	0.00	0.00	40.00	0.00	80.7	NaN	NaN	[74]
148	3	40.00	10.00	0.00	0.00	0.00	50.00	0.00	91.5	NaN	NaN	[74]
149	3	32.00	8.00	0.00	0.00	0.00	60.00	0.00	99.1	NaN	NaN	[74]
150	3	16.00	4.00	0.00	0.00	0.00	80.00	0.00	105.8	NaN	NaN	[74]
151	3	63.00	27.00	0.00	0.00	0.00	10.00	0.00	46.5	NaN	NaN	[74]
152	3	56.00	24.00	0.00	0.00	0.00	20.00	0.00	60.8	NaN	NaN	[74]
153	3	49.00	21.00	0.00	0.00	0.00	30.00	0.00	74.2	NaN	NaN	[74]
154	3	42.00	18.00	0.00	0.00	0.00	40.00	0.00	85.5	NaN	NaN	[74]
155	3	35.00	15.00	0.00	0.00	0.00	50.00	0.00	94.7	NaN	NaN	[74]
156	3	54.00	36.00	0.00	0.00	0.00	10.00	0.00	55	NaN	NaN	[74]
157	3	48.00	32.00	0.00	0.00	0.00	20.00	0.00	68.5	NaN	NaN	[74]
158	3	42.00	28.00	0.00	0.00	0.00	30.00	0.00	80.6	NaN	NaN	[74]
159	3	36.00	24.00	0.00	0.00	0.00	40.00	0.00	90.4	NaN	NaN	[74]
160	3	30.00	20.00	0.00	0.00	0.00	50.00	0.00	97.9	NaN	NaN	[74]
161	3	24.00	16.00	0.00	0.00	0.00	60.00	0.00	102.7	NaN	NaN	[74]
162	3	12.00	8.00	0.00	0.00	0.00	80.00	0.00	106.6	NaN	NaN	[74]
163	3	45.00	45.00	0.00	0.00	0.00	10.00	0.00	63.8	NaN	NaN	[74]
164	3	40.00	40.00	0.00	0.00	0.00	20.00	0.00	75.8	NaN	NaN	[74]
165	3	35.00	35.00	0.00	0.00	0.00	30.00	0.00	86.4	NaN	NaN	[74]
166	3	30.00	30.00	0.00	0.00	0.00	40.00	0.00	94.5	NaN	NaN	[74]
167	3	36.00	54.00	0.00	0.00	0.00	10.00	0.00	72.6	NaN	NaN	[74]
168	3	32.00	48.00	0.00	0.00	0.00	20.00	0.00	83.5	NaN	NaN	[74]
169	3	28.00	42.00	0.00	0.00	0.00	30.00	0.00	92	NaN	NaN	[74]
170	3	24.00	36.00	0.00	0.00	0.00	40.00	0.00	98.9	NaN	NaN	[74]
171	3	16.00	24.00	0.00	0.00	0.00	60.00	0.00	105.5	NaN	NaN	[74]
172	3	8.00	12.00	0.00	0.00	0.00	80.00	0.00	107.6	NaN	NaN	[74]
173	3	27.00	63.00	0.00	0.00	0.00	10.00	0.00	80.9	NaN	NaN	[74]
174	3	24.00	56.00	0.00	0.00	0.00	20.00	0.00	90.3	NaN	NaN	[74]
175	3	21.00	49.00	0.00	0.00	0.00	30.00	0.00	97.4	NaN	NaN	[74]
176	3	18.00	72.00	0.00	0.00	0.00	10.00	0.00	89.5	NaN	NaN	[74]
177	3	16.00	64.00	0.00	0.00	0.00	20.00	0.00	97	NaN	NaN	[74]
178	3	12.00	48.00	0.00	0.00	0.00	40.00	0.00	105.7	NaN	NaN	[74]
179	3	8.00	32.00	0.00	0.00	0.00	60.00	0.00	107.7	NaN	NaN	[74]
180	3	4.00	16.00	0.00	0.00	0.00	80.00	0.00	108.3	NaN	NaN	[74]
181	3	9.50	85.50	0.00	0.00	0.00	5.00	0.00	94.1	NaN	NaN	[74]
182	3	9.00	81.00	0.00	0.00	0.00	10.00	0.00	97.6	NaN	NaN	[74]
183	3	8.00	72.00	0.00	0.00	0.00	20.00	0.00	103.6	NaN	NaN	[74]
184	2	0.00	90.00	0.00	0.00	0.00	10.00	0.00	106.8	99.9	6.9	[74]
185	2	0.00	80.00	0.00	0.00	0.00	20.00	0.00	109.4	99.1	10.3	[74]
186	2	0.00	60.00	0.00	0.00	0.00	40.00	0.00	110.2	95.9	14.3	[74]
187	2	0.00	40.00	0.00	0.00	0.00	60.00	0.00	109.6	94.2	15.4	[74]
188	2	0.00	20.00	0.00	0.00	0.00	80.00	0.00	109	92.6	16.4	[74]
189	5	7.00	53.00	14.00	14.00	12.00	0.00	0.00	93.8	89.5	4.3	[112]
190	3	19.31	36.58	0.00	0.00	44.10	0.00	0.00	92	84.3	7.7	[113]
191	4	9.97	31.33	21.24	0.00	37.45	0.00	0.00	90.9	82.7	8.2	[113]
192	2	9.00	91.00	0.00	0.00	0.00	0.00	0.00	91	91	0	[114]
193	3	12.50	72.50	0.00	0.00	15.00	0.00	0.00	90.5	88.0	2.5	[114]
194	3	17.50	52.50	0.00	0.00	30.00	0.00	0.00	89.5	84.7	4.8	[114]
195	3	5.10	84.70	10.20	0.00	0.00	0.00	0.00	92.9	90.2	2.7	[114]
196	3	3.00	77.00	20.00	0.00	0.00	0.00	0.00	93	88.5	4.5	[114]
197	4	10.00	65.00	10.00	0.00	15.00	0.00	0.00	91.2	86.8	4.4	[114]
198	4	7.00	58.00	20.00	0.00	15.00	0.00	0.00	91.7	85.2	6.5	[114]
199	4	13.86	46.53	7.67	0.00	31.93	0.00	0.00	91.4	84.9	6.5	[114]
200	4	11.00	39.00	20.00	0.00	30.00	0.00	0.00	90.9	82.7	8.2	[114]
201	2	30.00	70.00	0.00	0.00	0.00	0.00	0.00	70	70	0	[114]
202	3	33.00	52.00	0.00	0.00	15.00	0.00	0.00	71.2	69	2.2	[114]
203	3	40.00	30.00	0.00	0.00	30.00	0.00	0.00	68.4	63.7	4.7	[114]
204	3	25.00	65.00	10.00	0.00	0.00	0.00	0.00	74.2	72.6	1.6	[114]

Continued on next page

id	#	P	I	O	N	A	Ox	Ether	RON	MON	S	Ref
205	3	23.00	57.00	20.00	0.00	0.00	0.00	0.00	74.6	72	2.6	[114]
206	4	31.00	44.00	10.00	0.00	15.00	0.00	0.00	72	68	4	[114]
207	4	29.00	36.00	20.00	0.00	15.00	0.00	0.00	72	67.2	4.8	[114]
208	2	0.00	0.00	0.00	90.00	0.00	10.00	0.00	101	85.9	15.1	[98]
209	2	0.00	0.00	0.00	75.00	0.00	25.00	0.00	102.3	86.3	16	[98]
210	2	0.00	0.00	0.00	60.00	0.00	40.00	0.00	103.3	86.6	16.7	[98]
211	2	0.00	0.00	0.00	40.00	0.00	60.00	0.00	104.8	87.2	17.6	[98]
212	2	0.00	0.00	90.00	0.00	0.00	10.00	0.00	81	68.5	12.5	[98]
213	2	0.00	0.00	75.00	0.00	0.00	25.00	0.00	89.2	74.5	14.7	[98]
214	2	0.00	0.00	60.00	0.00	0.00	40.00	0.00	96.5	79.5	17	[98]
215	2	0.00	0.00	40.00	0.00	0.00	60.00	0.00	101.7	84	17.7	[98]
216	5	11.60	54.90	4.70	0.00	18.90	9.90	0.00	92.1	NaN	NaN	[115]
217	2	50.00	0.00	0.00	0.00	50.00	0.00	0.00	64.1	58.3	5.8	[116]
218	2	50.00	0.00	50.00	0.00	0.00	0.00	0.00	40.3	31.1	9.2	[116]
219	2	0.00	50.00	0.00	0.00	50.00	0.00	0.00	110.5	99.5	11	[116]
220	2	0.00	50.00	50.00	0.00	0.00	0.00	0.00	88.2	81.2	7	[116]
221	2	0.00	0.00	50.00	0.00	50.00	0.00	0.00	92.5	77.8	14.7	[116]
222	3	16.09	10.30	0.00	0.00	73.60	0.00	0.00	99.8	88.7	11.1	[117]
223	4	16.22	38.60	0.00	0.00	21.66	23.52	0.00	101.6	90.9	10.7	[117]
224	4	24.43	11.22	0.00	0.00	24.93	39.41	0.00	99.2	87	12.2	[117]
225	3	21.21	50.47	0.00	0.00	28.32	0.00	0.00	87	81.8	5.2	[117]
226	3	19.00	76.00	0.00	0.00	0.00	0.00	5.00	83.1	NaN	NaN	[118]
227	3	18.00	72.00	0.00	0.00	0.00	0.00	10.00	86.3	NaN	NaN	[118]
228	3	17.00	68.00	0.00	0.00	0.00	0.00	15.00	89.1	NaN	NaN	[118]
229	3	19.00	76.00	0.00	0.00	0.00	5.00	0.00	84.9	NaN	NaN	[118]
230	3	18.00	72.00	0.00	0.00	0.00	10.00	0.00	89.7	NaN	NaN	[118]
231	3	17.00	68.00	0.00	0.00	0.00	15.00	0.00	93.7	NaN	NaN	[118]
232	3	9.50	85.50	0.00	0.00	0.00	0.00	5.00	92.5	NaN	NaN	[118]
233	3	9.00	81.00	0.00	0.00	0.00	0.00	10.00	94.9	NaN	NaN	[118]
234	3	8.50	76.50	0.00	0.00	0.00	0.00	15.00	97.1	NaN	NaN	[118]
235	3	9.50	85.50	0.00	0.00	0.00	5.00	0.00	94	NaN	NaN	[118]
236	3	9.00	81.00	0.00	0.00	0.00	10.00	0.00	97.8	NaN	NaN	[118]
237	3	8.50	76.50	0.00	0.00	0.00	15.00	0.00	100.8	NaN	NaN	[118]
238	2	0.00	95.00	0.00	0.00	0.00	0.00	5.00	102.1	NaN	NaN	[118]
239	2	0.00	90.00	0.00	0.00	0.00	0.00	10.00	104.2	NaN	NaN	[118]
240	2	0.00	85.00	0.00	0.00	0.00	0.00	15.00	106	NaN	NaN	[118]
241	2	0.00	95.00	0.00	0.00	0.00	5.00	0.00	103.9	NaN	NaN	[118]
242	4	10.20	37.80	0.00	0.00	12.00	40.00	0.00	98.75	NaN	NaN	[119]
243	3	18.00	62.00	0.00	0.00	0.00	20.00	0.00	92	NaN	NaN	[119]
244	3	17.00	69.00	0.00	0.00	14.00	0.00	0.00	87	NaN	NaN	[119]
245	3	16.67	16.67	0.00	0.00	66.67	0.00	0.00	98	87.4	10.6	[120]
246	2	0.00	50.00	0.00	0.00	50.00	0.00	0.00	110	99.3	10.7	[120]
247	2	50.00	0.00	0.00	0.00	50.00	0.00	0.00	65.9	57.7	8.2	[120]
248	3	33.33	33.33	0.00	0.00	33.33	0.00	0.00	76.2	70.9	5.3	[120]
249	3	16.67	66.67	0.00	0.00	16.67	0.00	0.00	87	84	3	[120]
250	3	66.67	16.67	0.00	0.00	16.67	0.00	0.00	39	37	2	[120]
251	2	50.00	0.00	0.00	0.00	50.00	0.00	0.00	65.1	58	7.1	[120]
252	2	42.00	0.00	0.00	0.00	58.00	0.00	0.00	75.6	66.9	8.7	[120]
253	2	34.00	0.00	0.00	0.00	66.00	0.00	0.00	85.2	74.8	10.4	[120]
254	3	11.00	15.00	0.00	0.00	74.00	0.00	0.00	103.3	92.6	10.7	[120]
255	3	6.00	20.00	0.00	0.00	74.00	0.00	0.00	107.6	96.6	11	[120]
256	2	0.00	26.00	0.00	0.00	74.00	0.00	0.00	113	100.8	12.2	[120]
257	2	30.00	0.00	0.00	0.00	70.00	0.00	0.00	89.3	78.2	11.1	[120]
258	2	26.00	0.00	0.00	0.00	74.00	0.00	0.00	93.4	81.5	11.9	[120]
259	3	21.00	5.00	0.00	0.00	74.00	0.00	0.00	96.9	85.2	11.7	[120]
260	3	16.00	10.00	0.00	0.00	74.00	0.00	0.00	99.8	88.7	11.1	[120]
261	2	35.00	0.00	0.00	0.00	65.00	0.00	0.00	83.9	73.2	10.7	[120]
262	2	36.00	0.00	0.00	0.00	64.00	0.00	0.00	82.3	73.1	9.2	[120]
263	2	38.00	0.00	0.00	0.00	62.00	0.00	0.00	80.5	70.3	10.2	[120]
264	2	50.00	0.00	0.00	0.00	50.00	0.00	0.00	64.1	58.1	6	[120]
265	2	25.00	0.00	0.00	0.00	75.00	0.00	0.00	94.2	82.6	11.6	[120]
266	3	17.00	63.00	0.00	0.00	20.00	0.00	0.00	88	85	3	[120]
267	3	17.00	69.00	0.00	0.00	14.00	0.00	0.00	87	85	2	[120]